

Яндекс

Яндекс

Эксплуатация Hadoop: tips and tricks

Игорь Андреев, системный администратор

Disclaimer

- Доклад не про программирование
- Доклад не является обзором Hadoop
- Сервисы находились в длительной эксплуатации
- Адаптация под инфраструктуру Яндекса

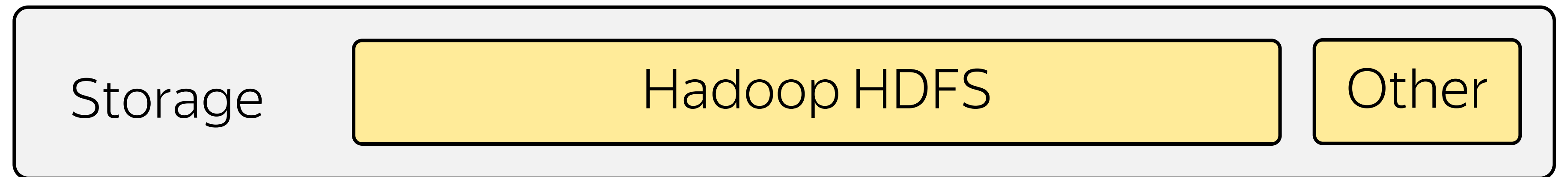
План доклада

- 1 | Краткий обзор Hadoop
- 2 | Примеры использования
- 3 | Система выкладки
- 4 | Проблемы и решения

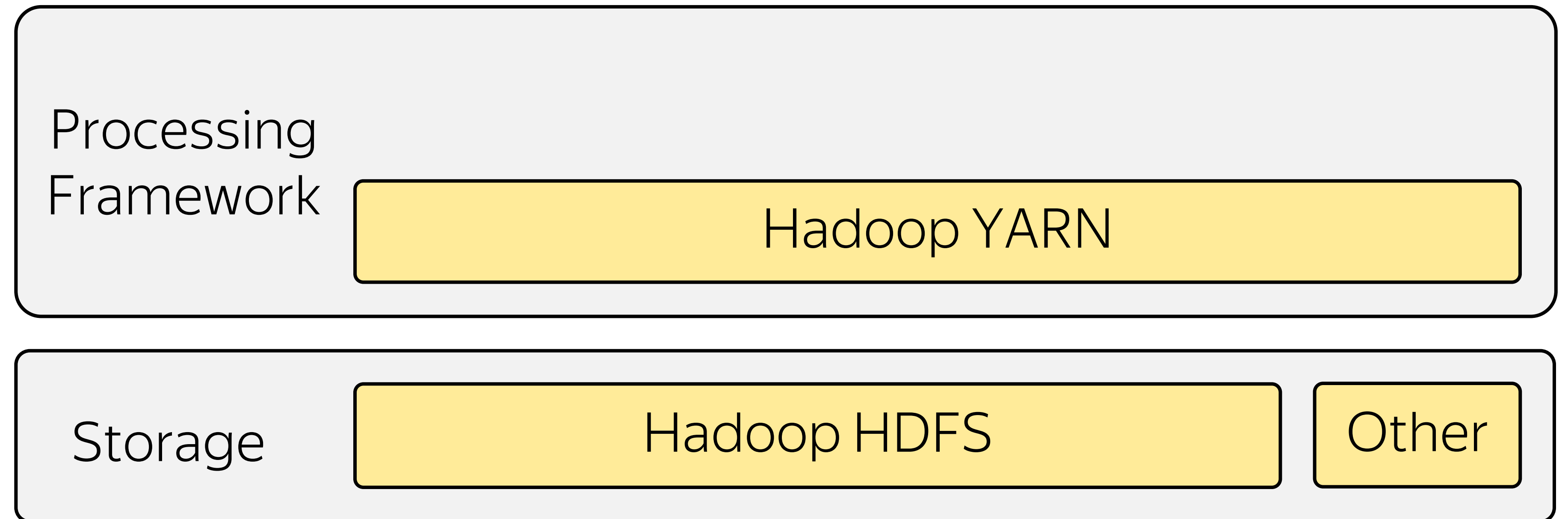
Краткий обзор Hadoop



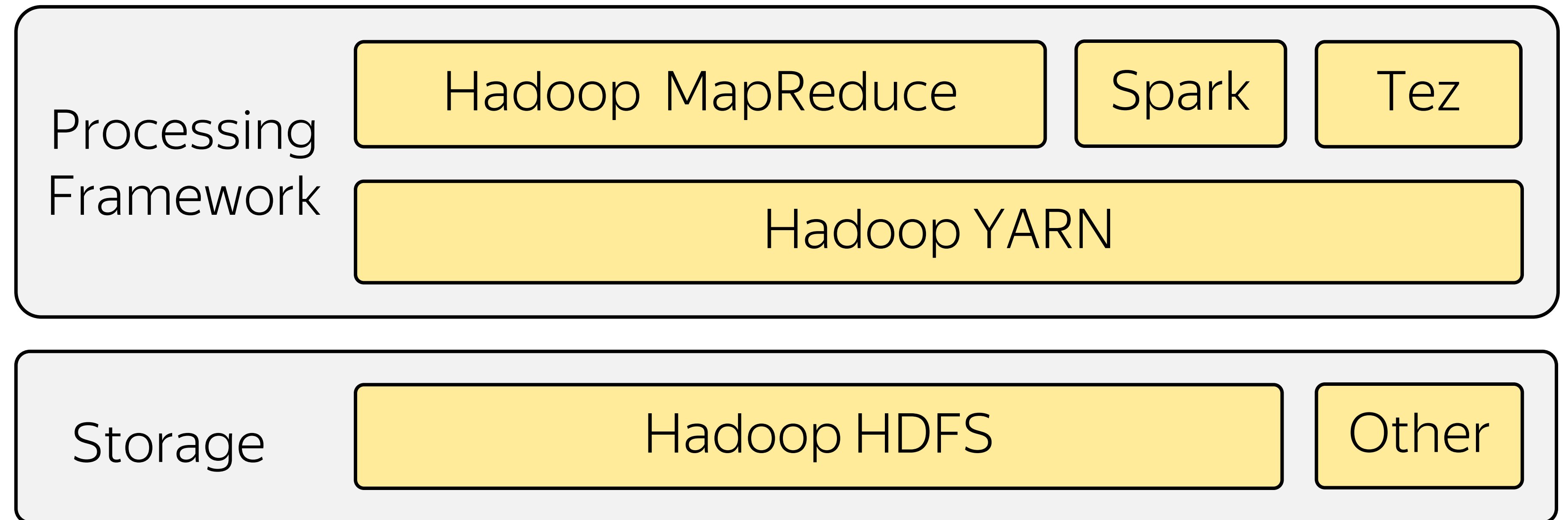
Hadoop Ecosystem



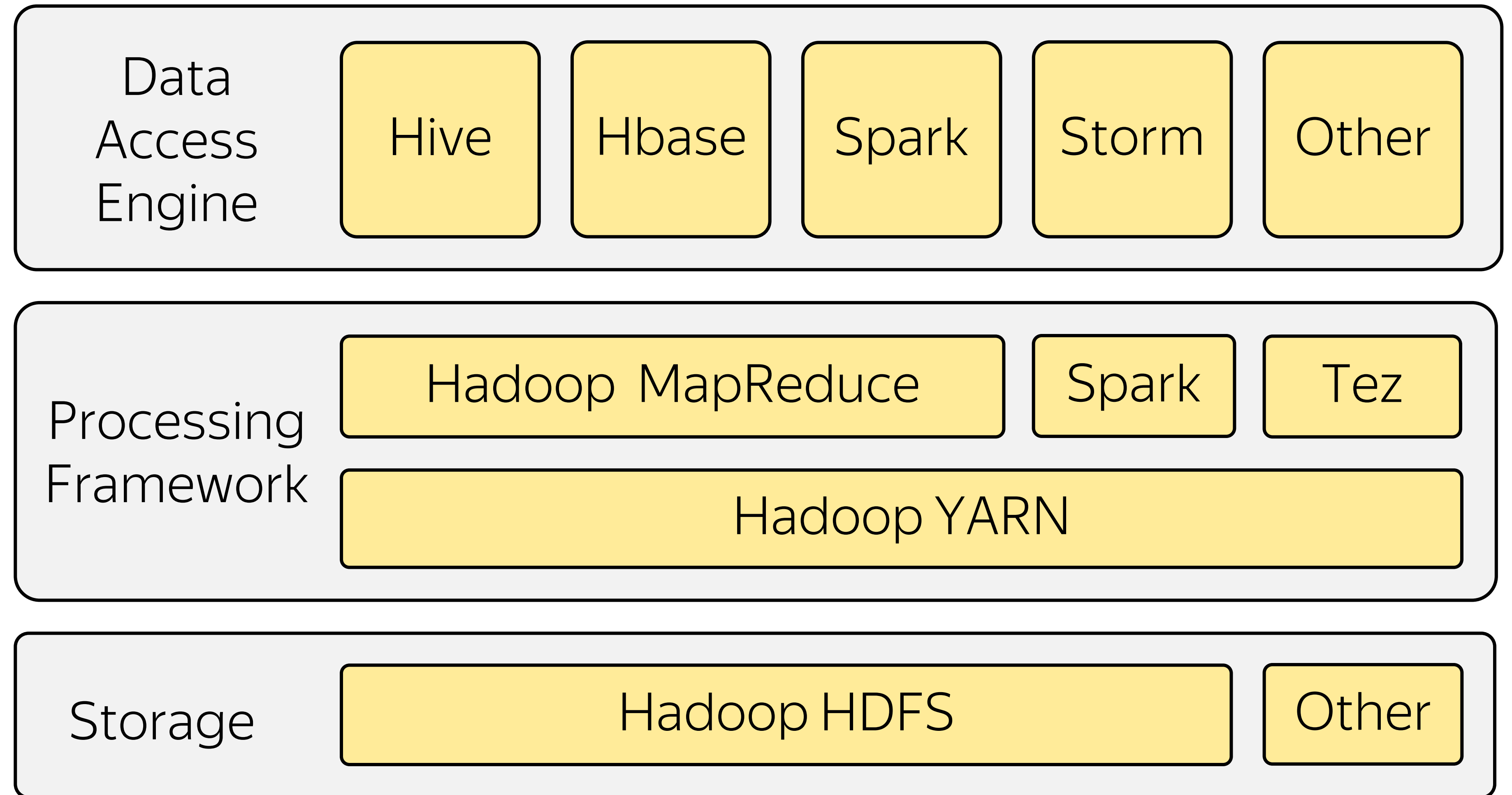
Hadoop Ecosystem



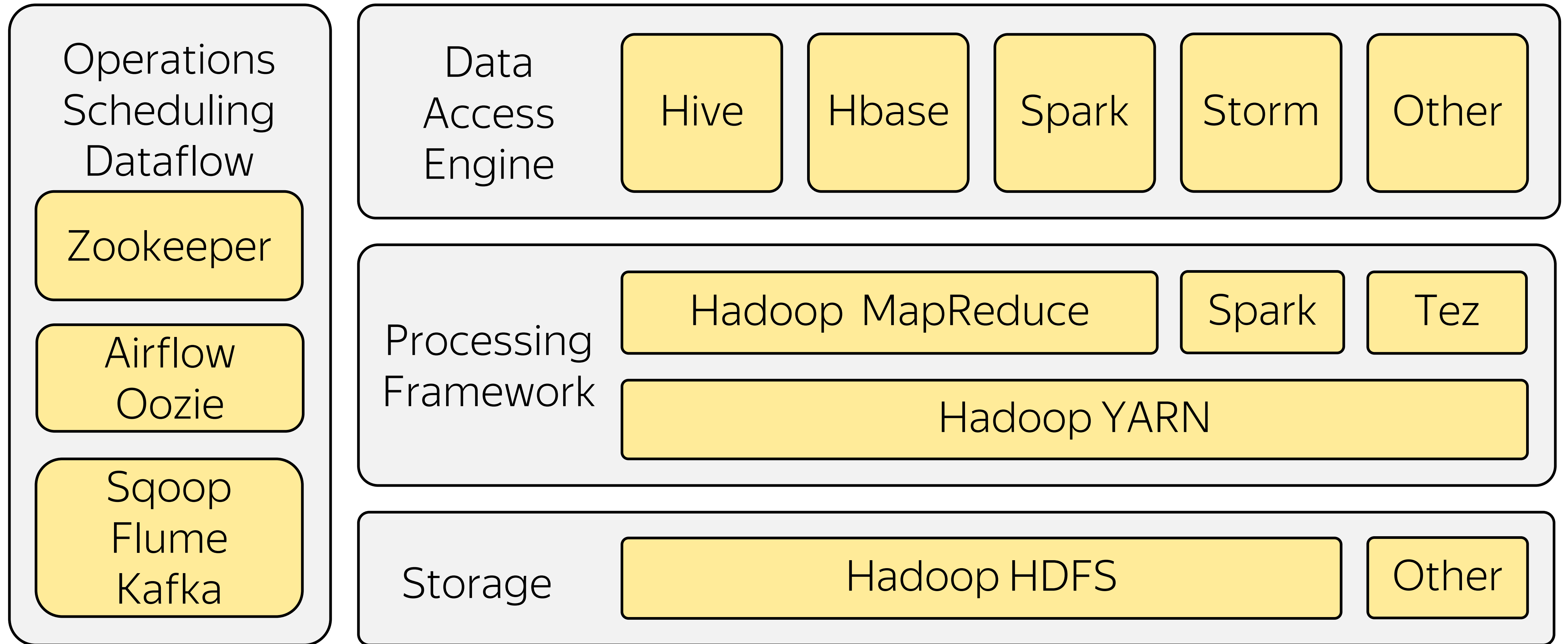
Hadoop Ecosystem



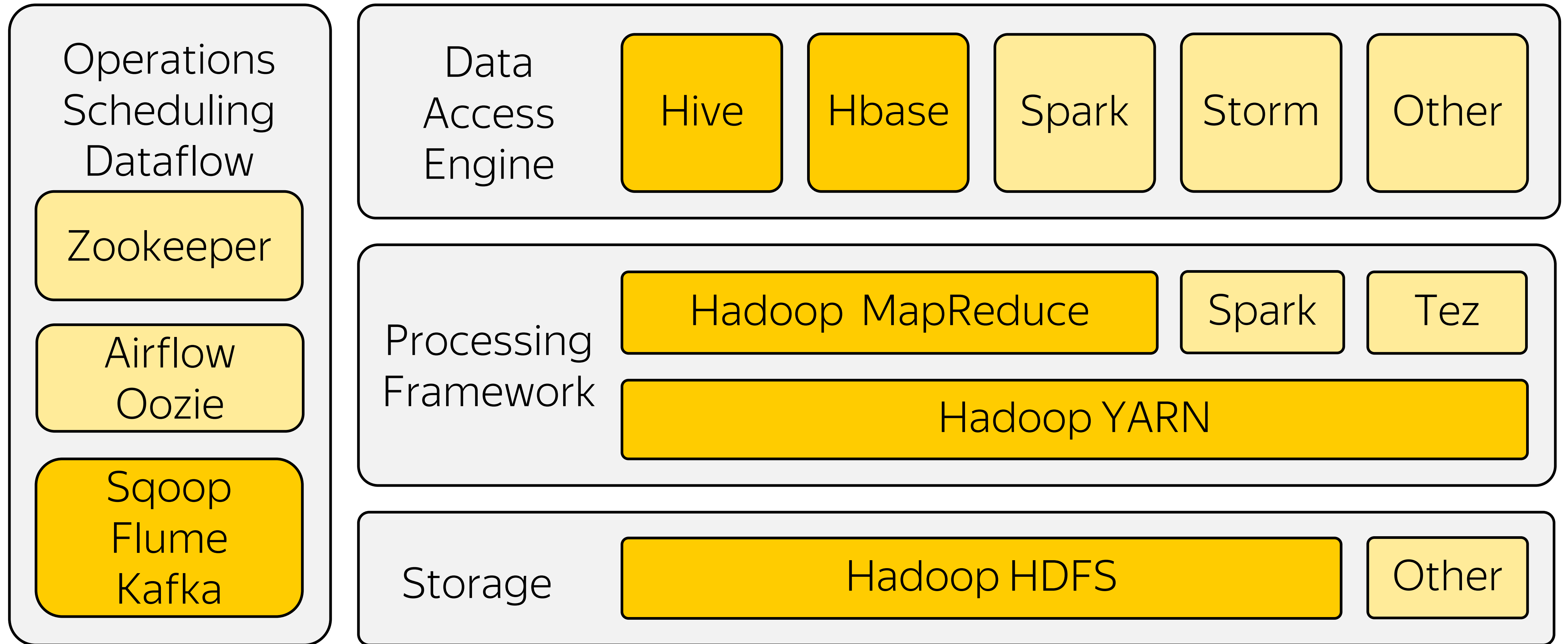
Hadoop Ecosystem



Hadoop Ecosystem



Hadoop Ecosystem

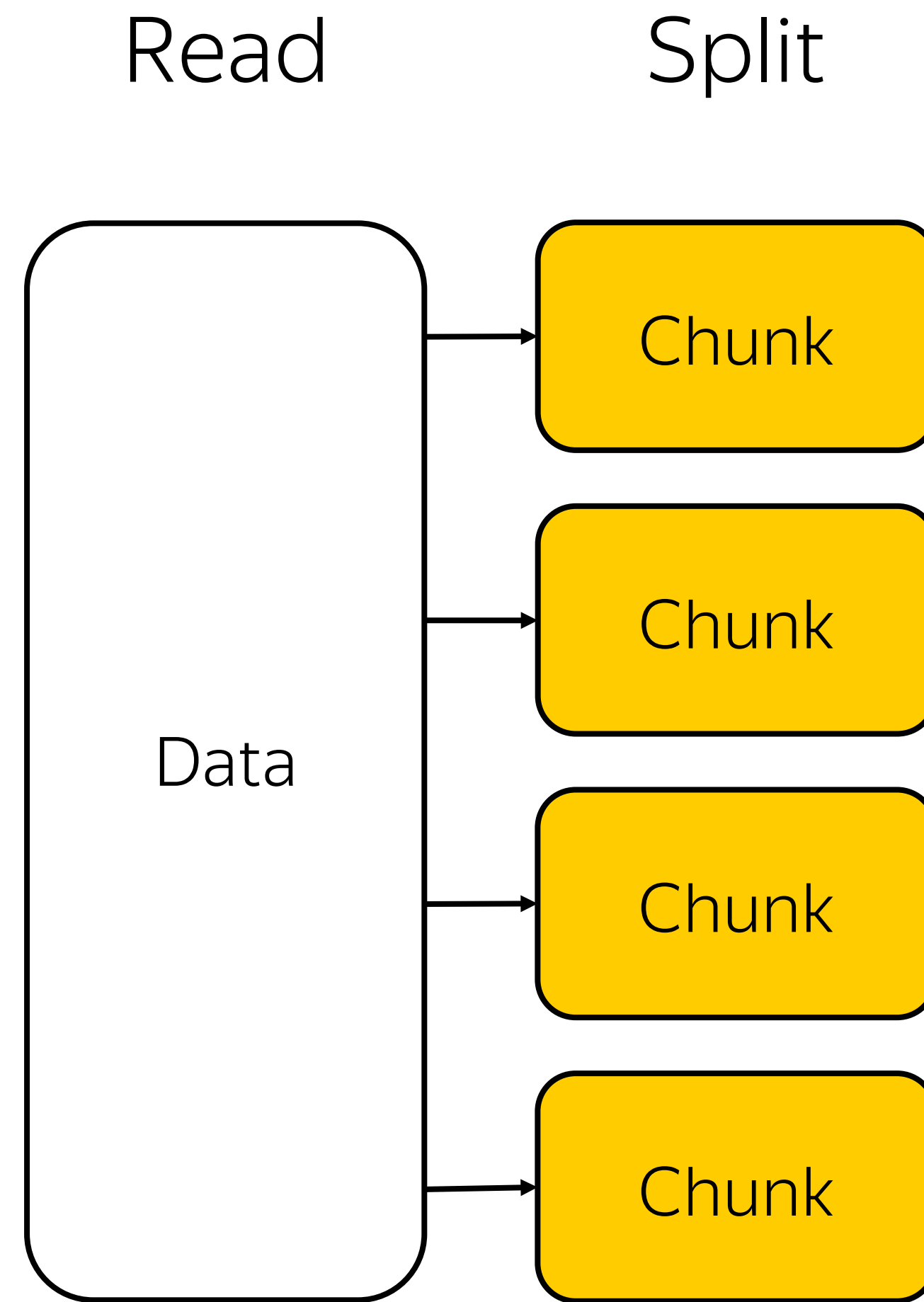


Map Reduce

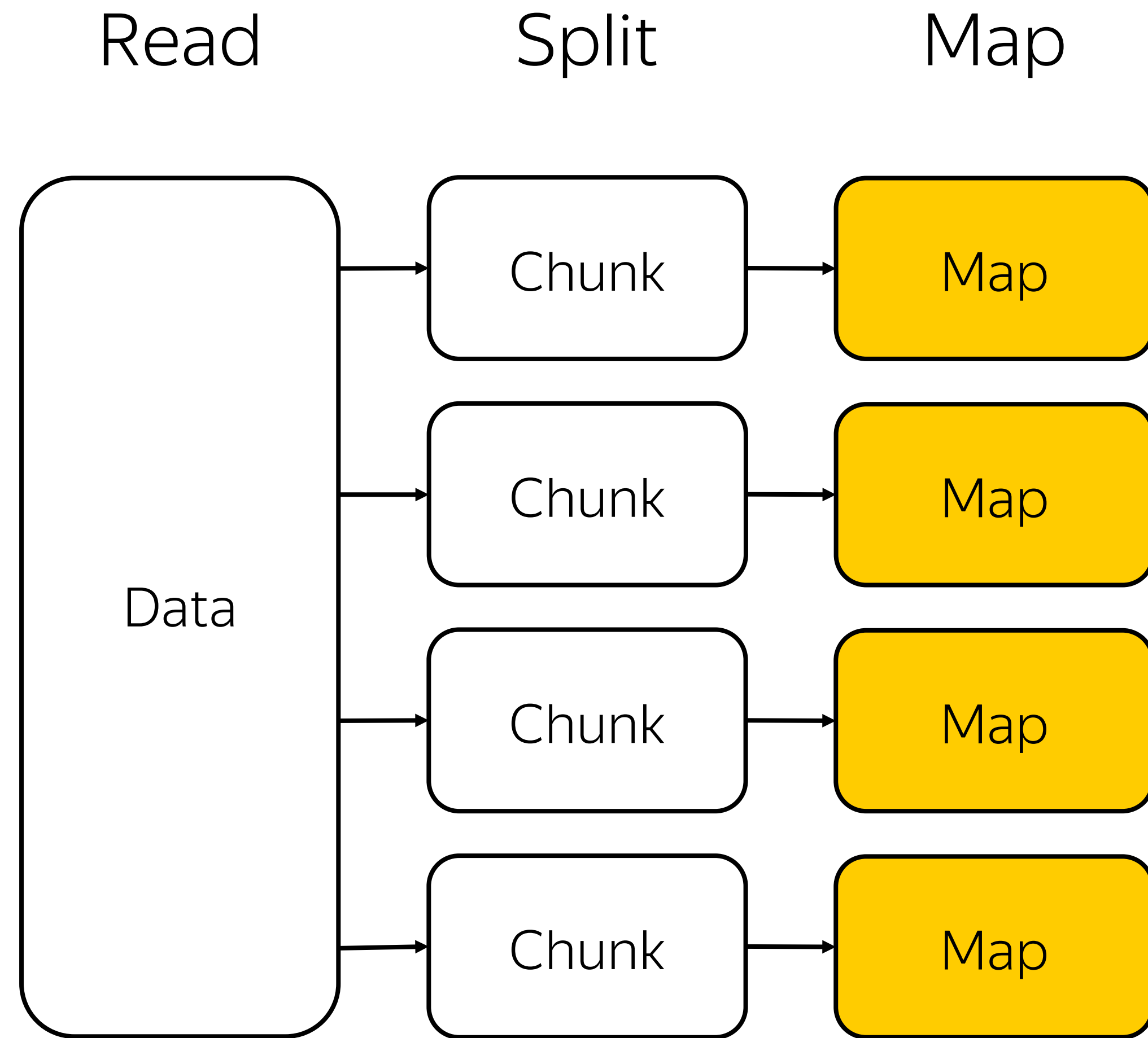
Read



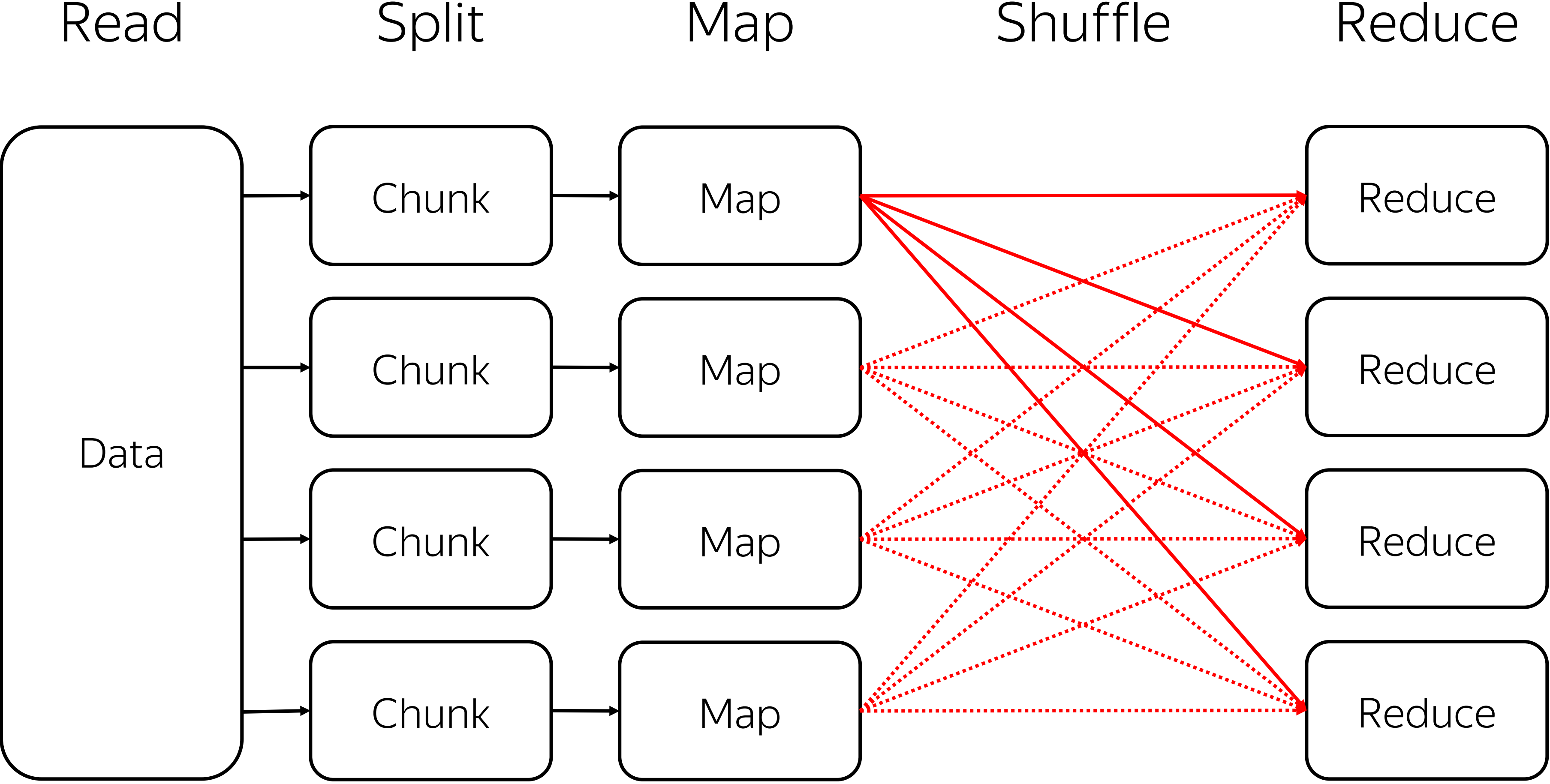
Map Reduce



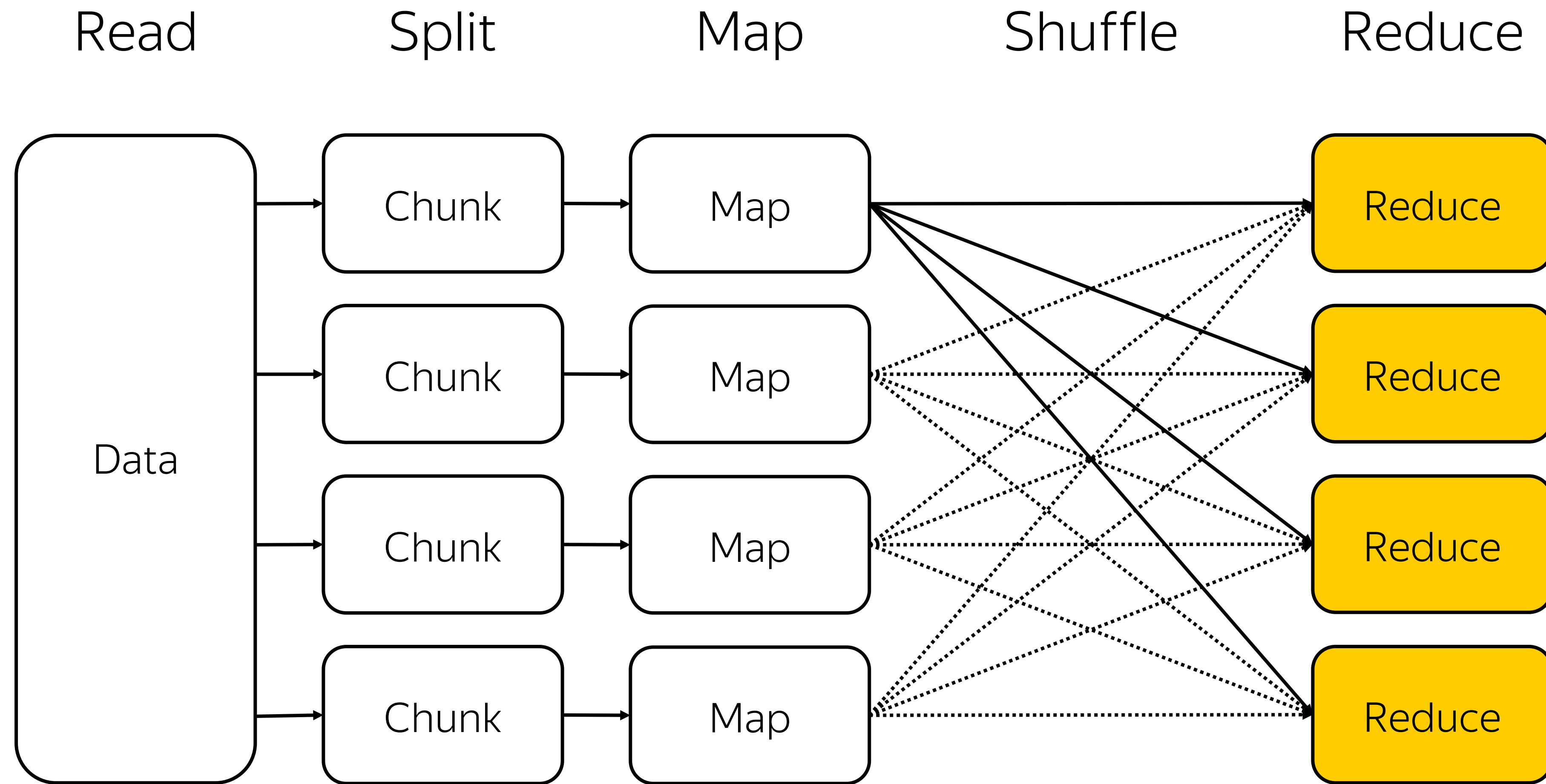
Map Reduce



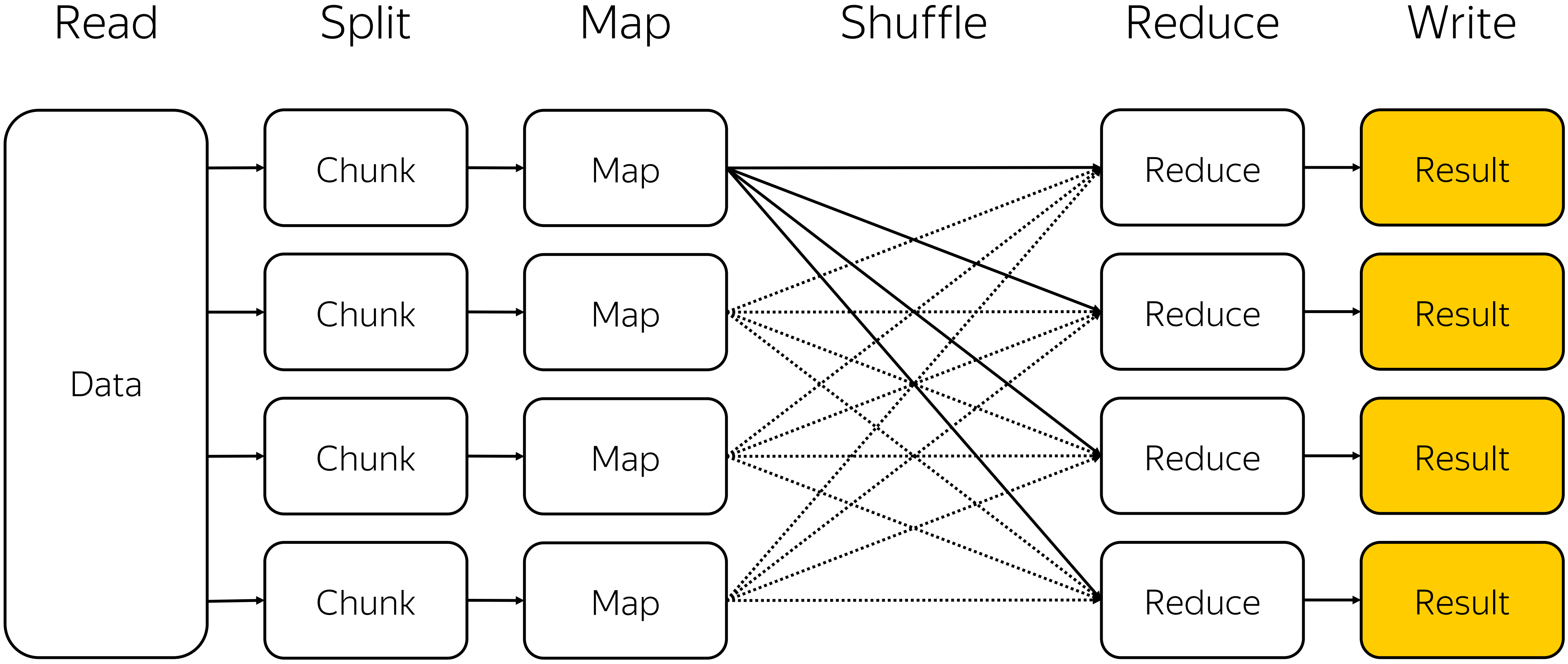
Map Reduce



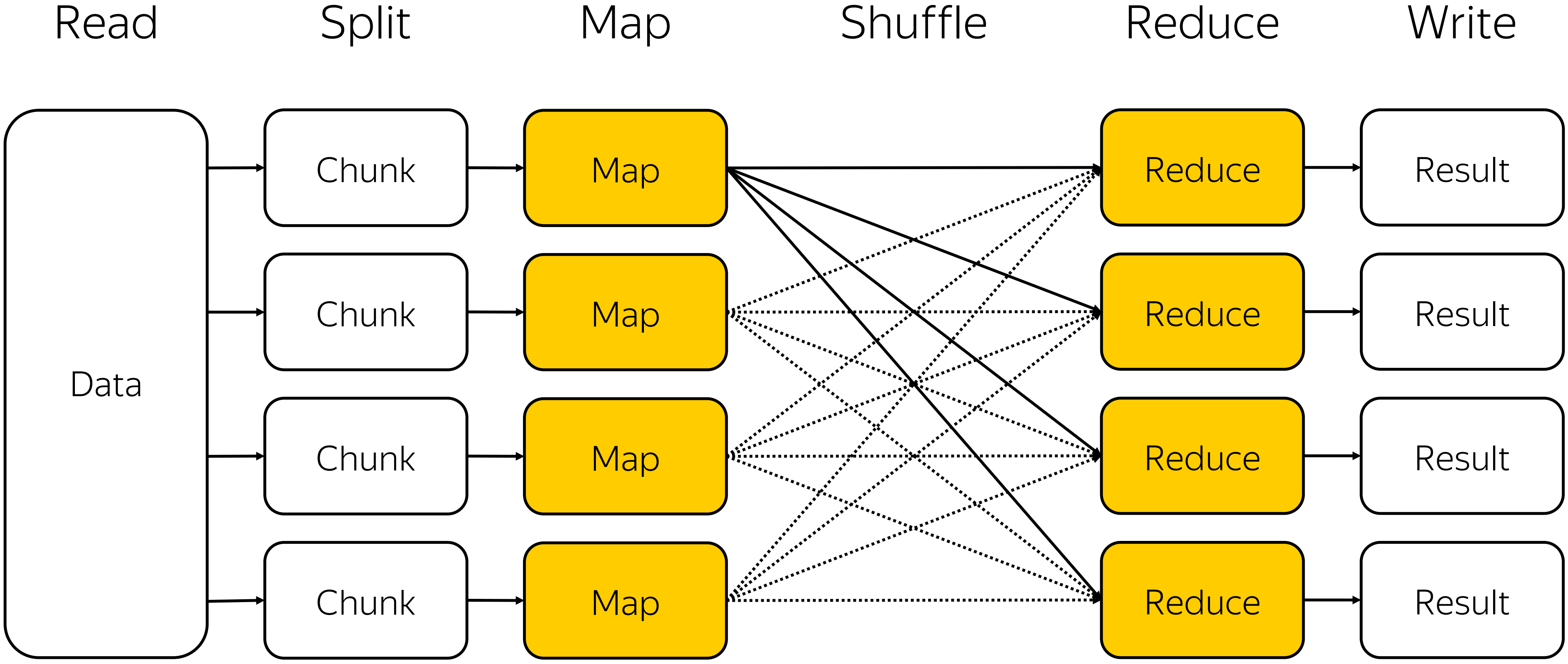
Map Reduce



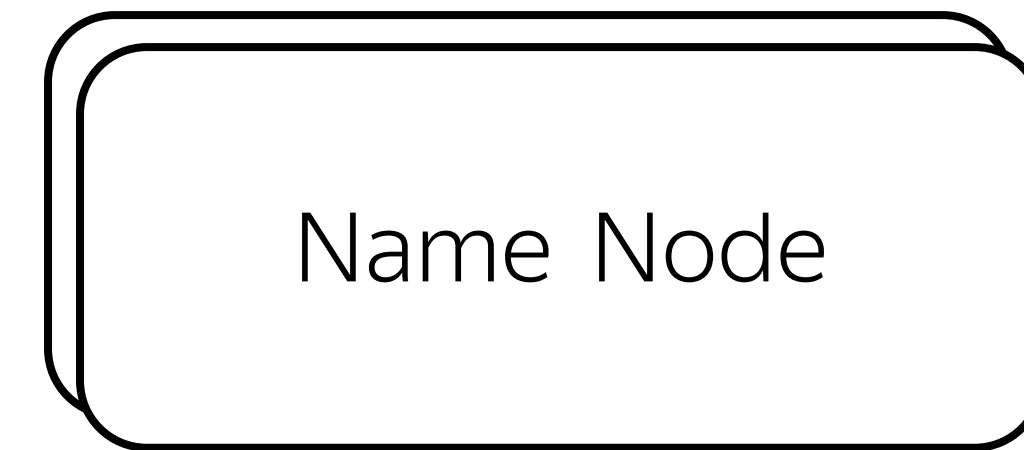
Map Reduce



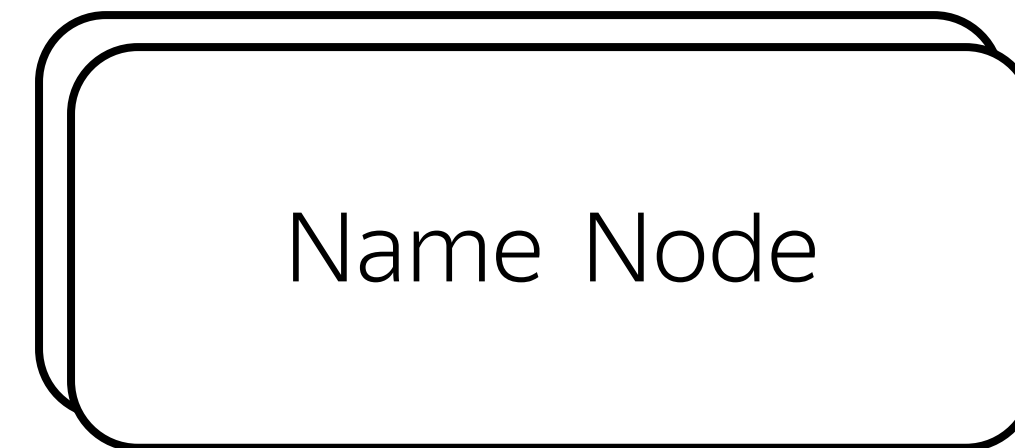
Map Reduce



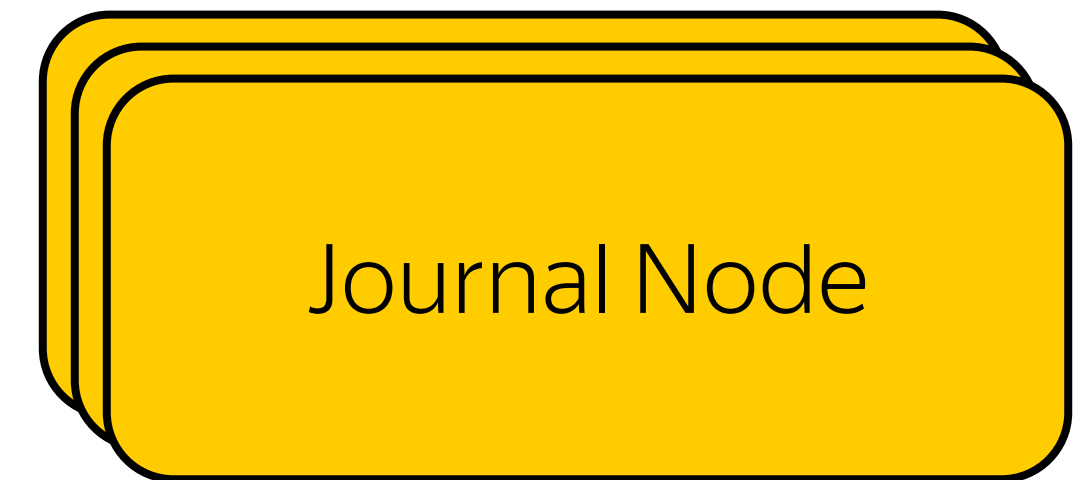
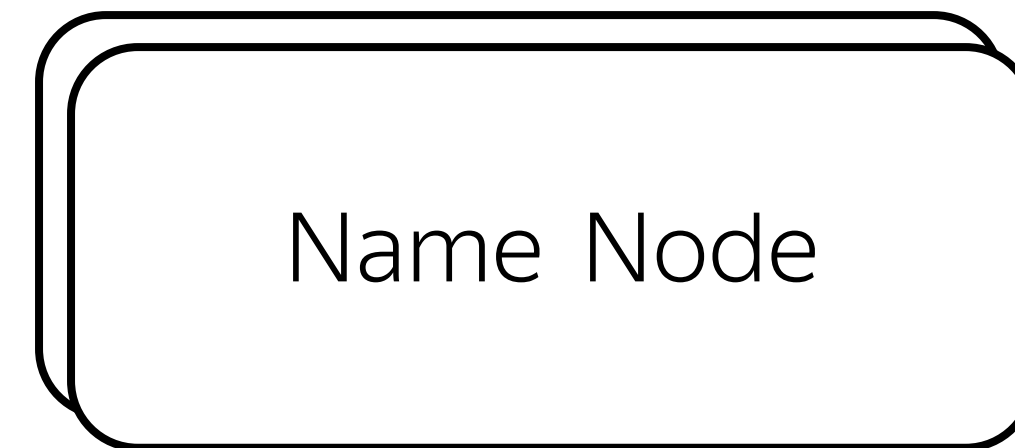
Раскладка сервисов по серверам: HDFS



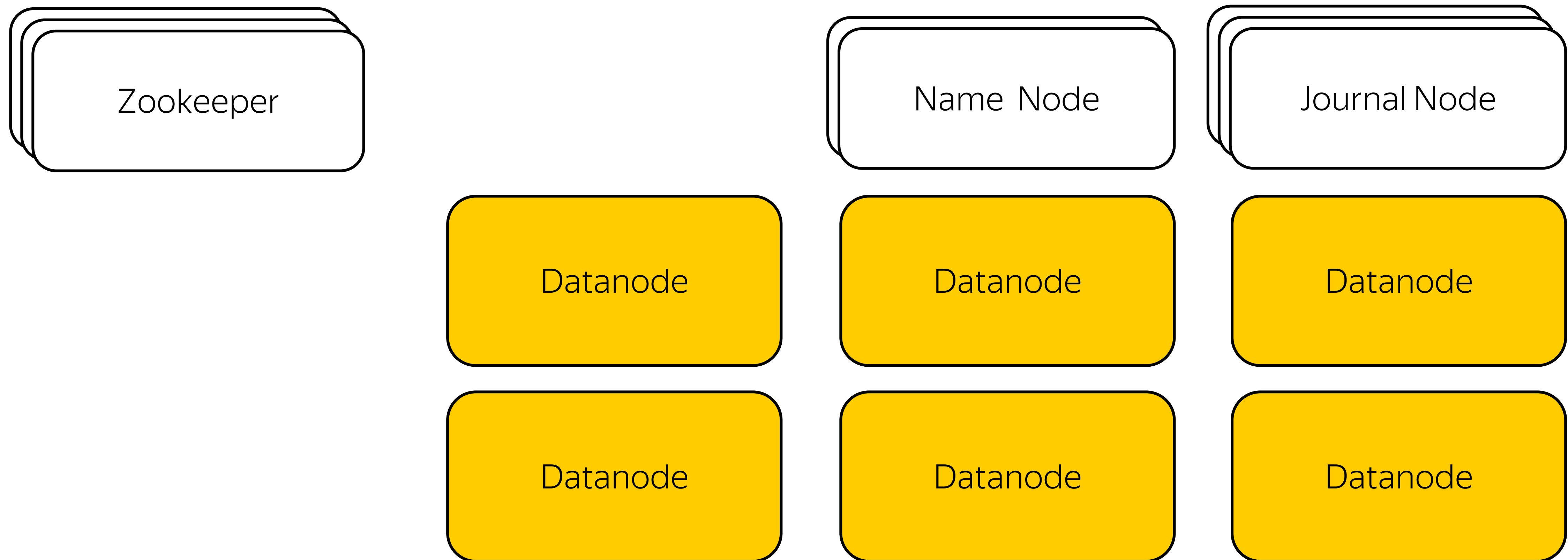
Раскладка сервисов по серверам: HDFS



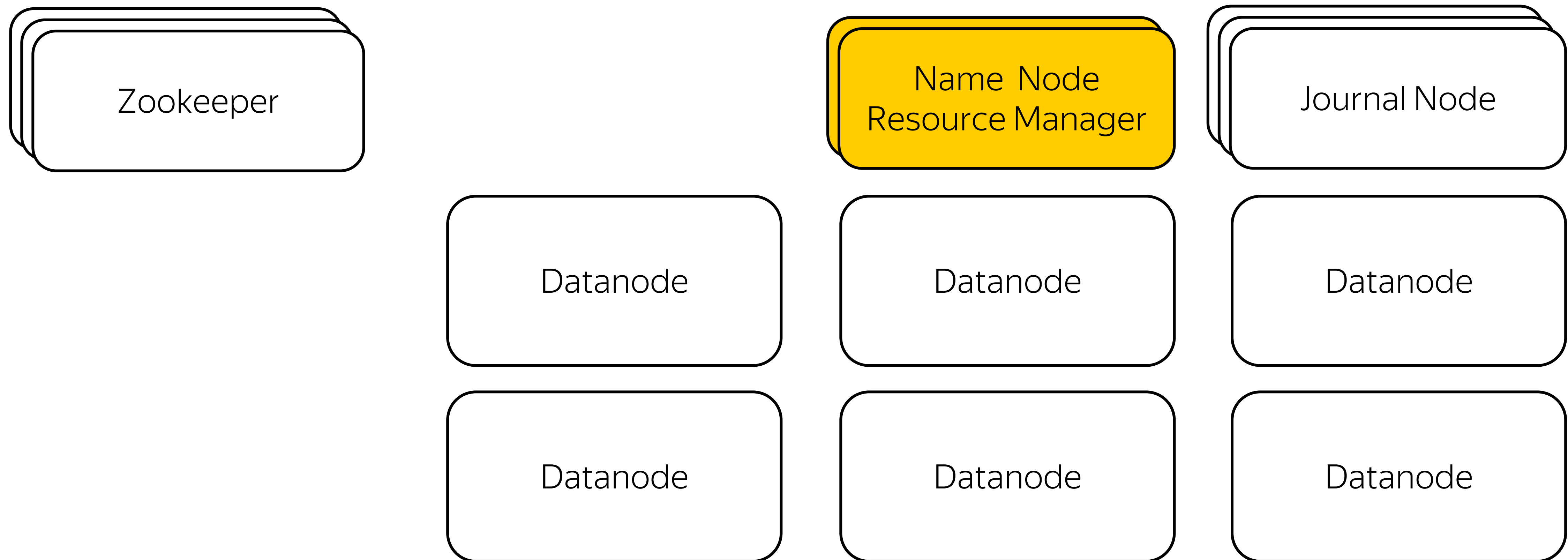
Раскладка сервисов по серверам: HDFS



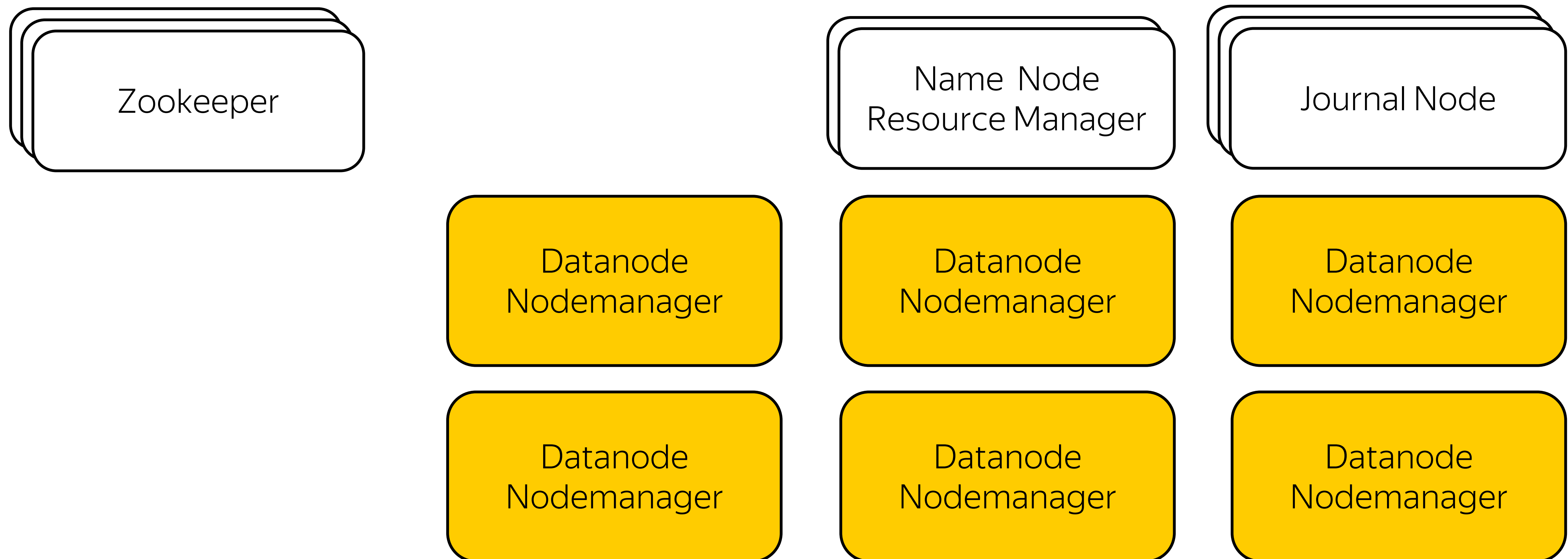
Раскладка сервисов по серверам: HDFS



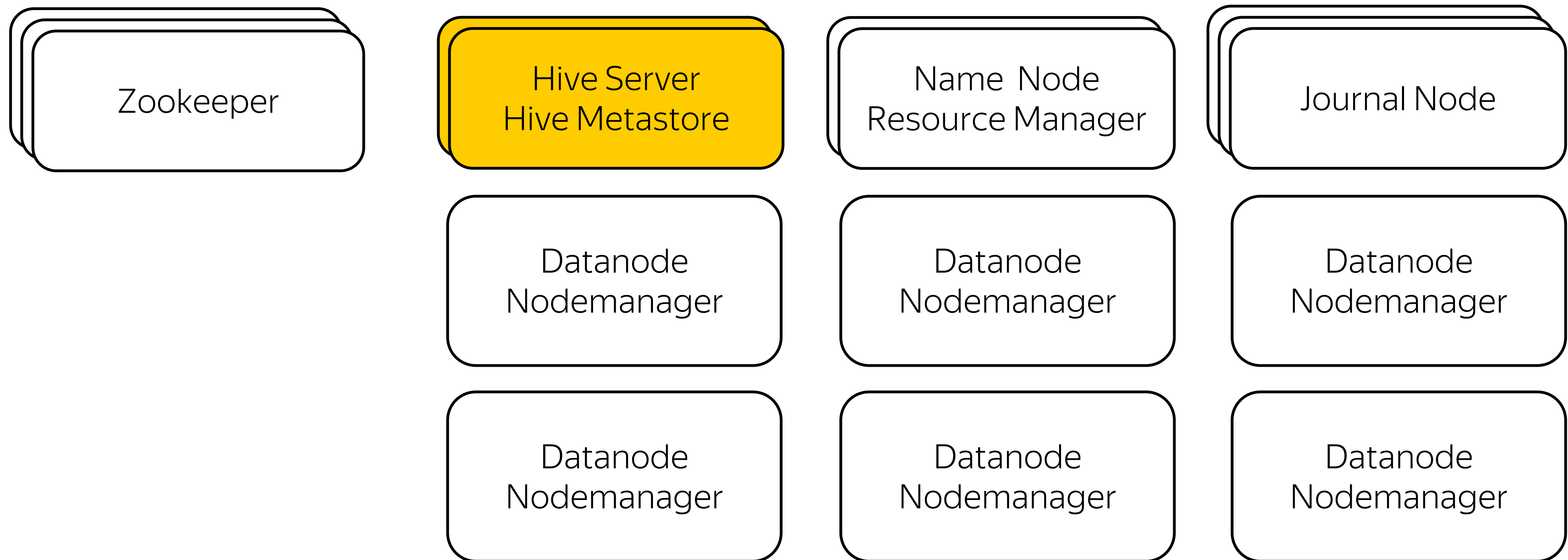
Раскладка сервисов по серверам: YARN



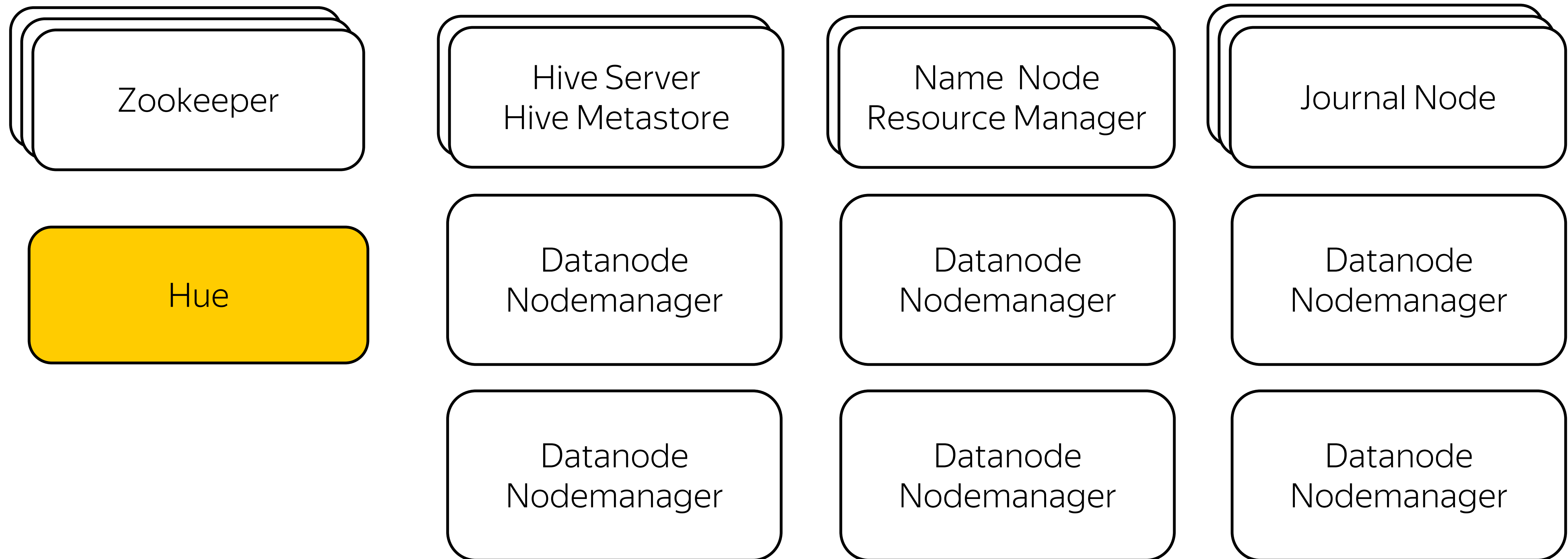
Раскладка сервисов по серверам: YARN



Раскладка сервисов по серверам: Hive



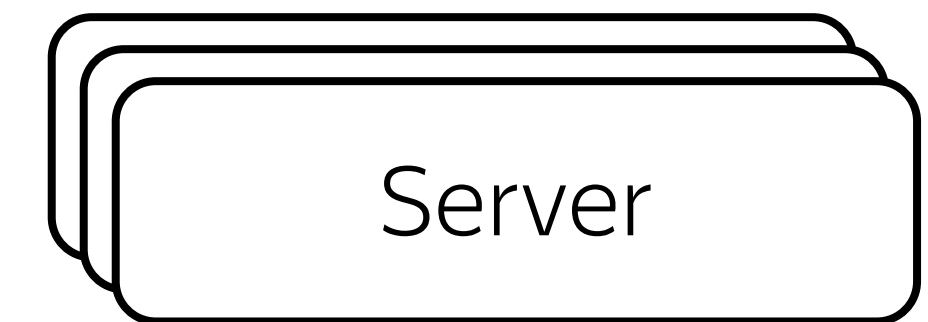
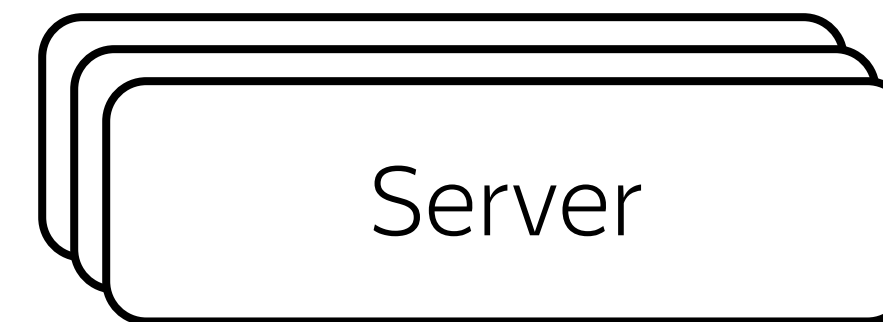
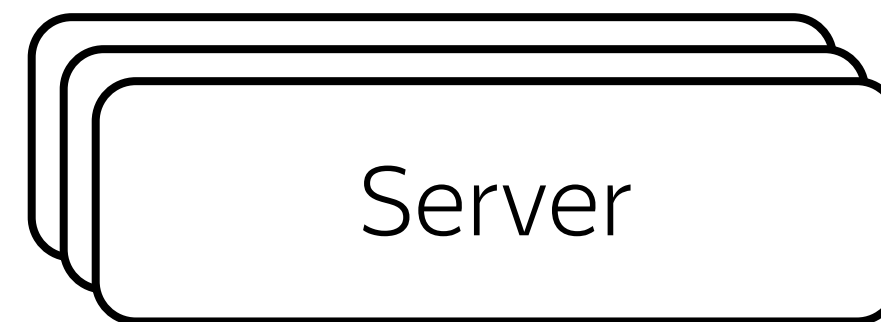
Раскладка сервисов по серверам: Hue



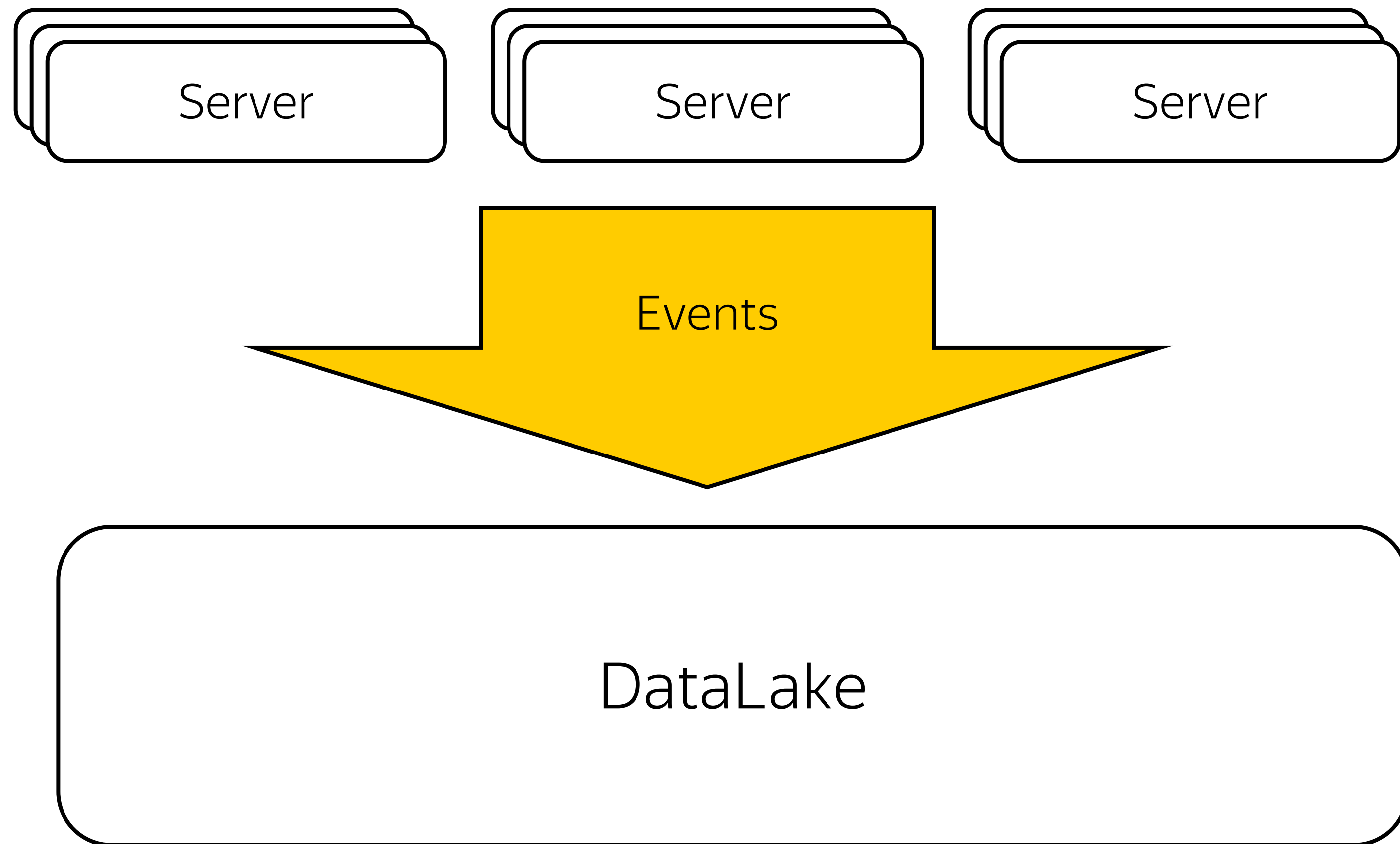
Примеры использования



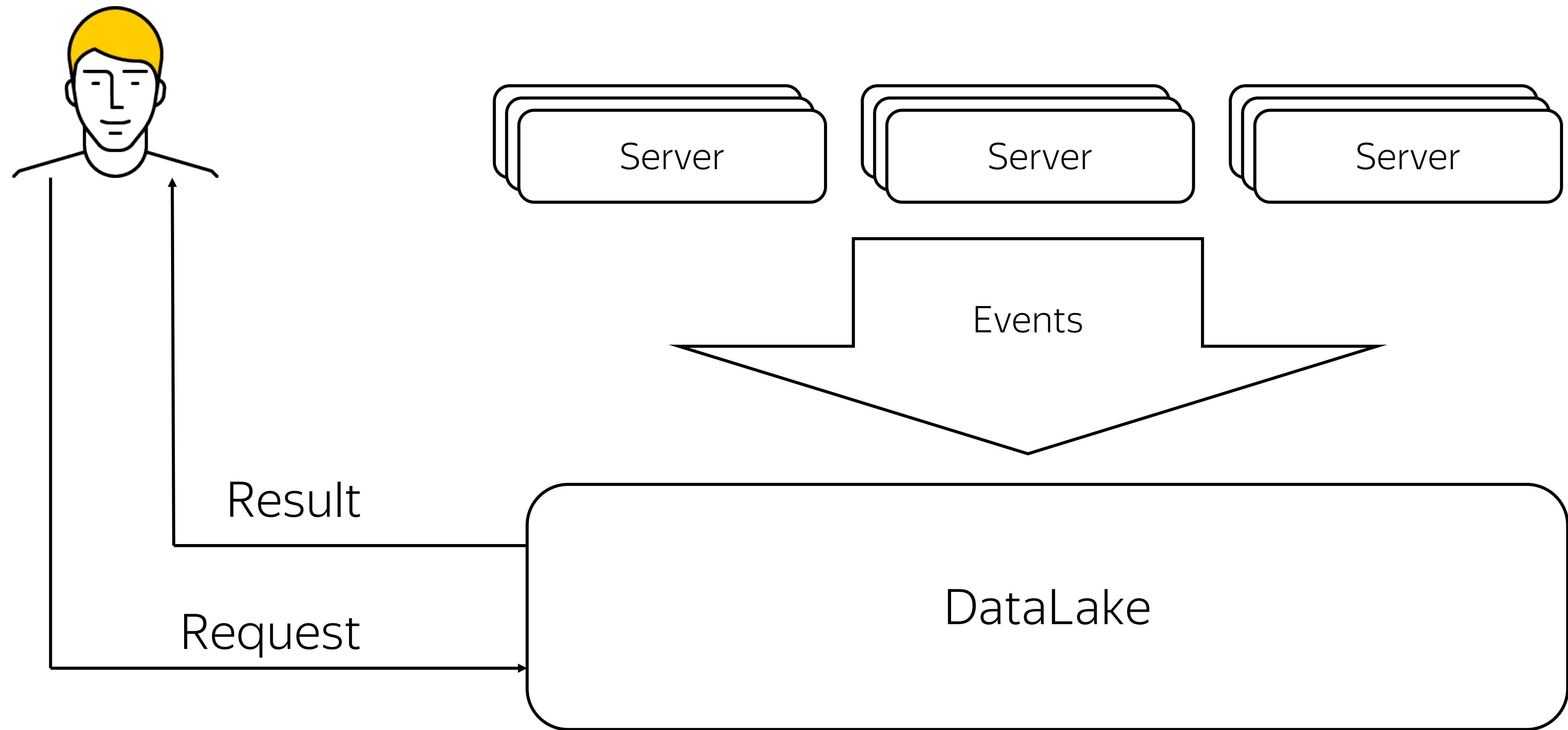
Поставка логов в DataLake



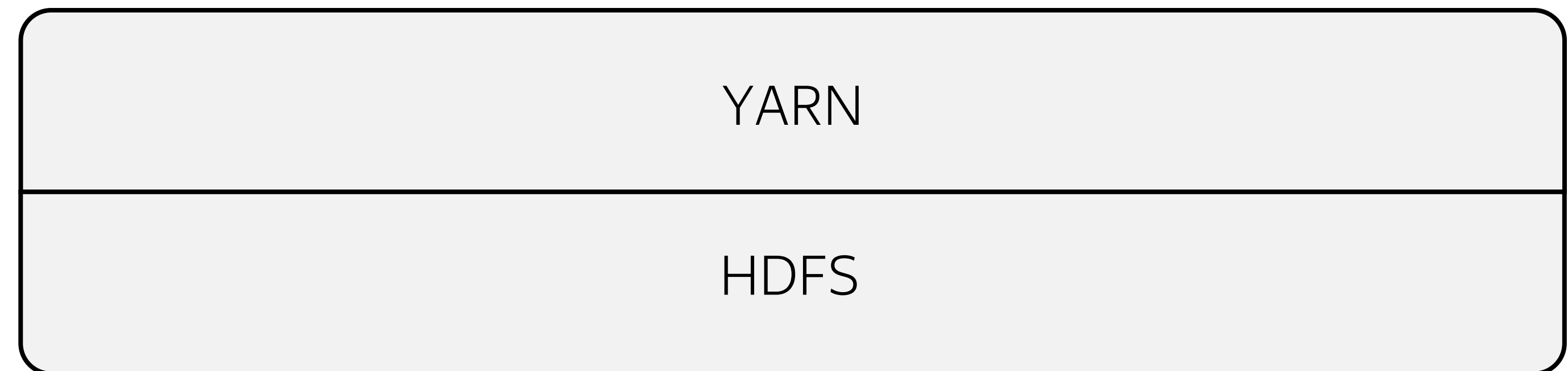
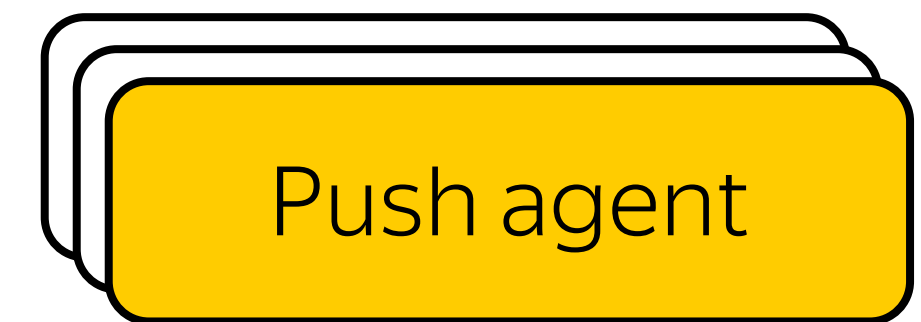
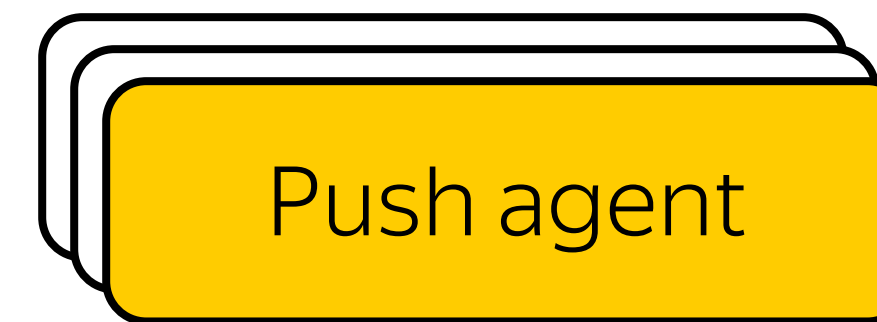
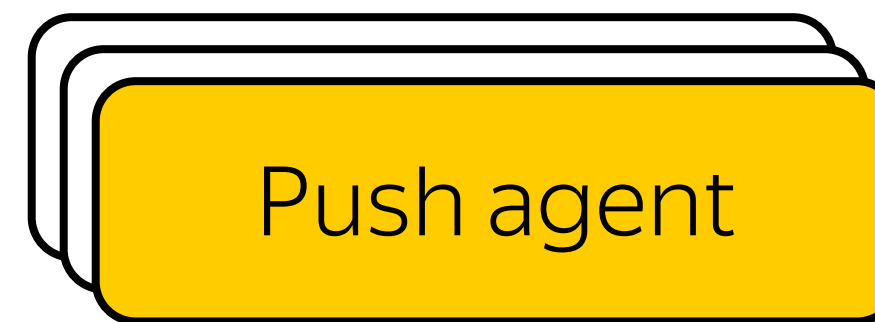
Поставка логов в DataLake



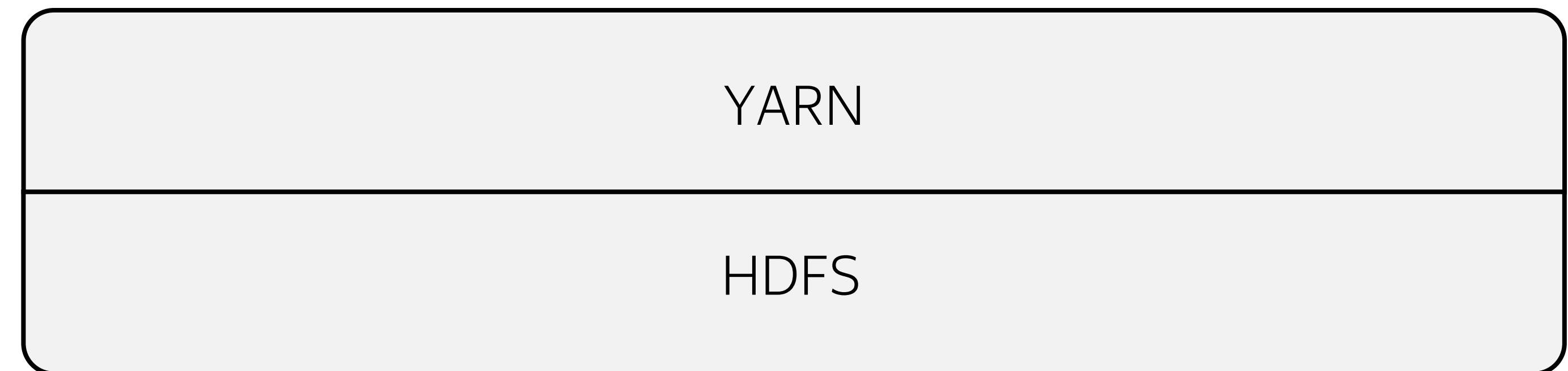
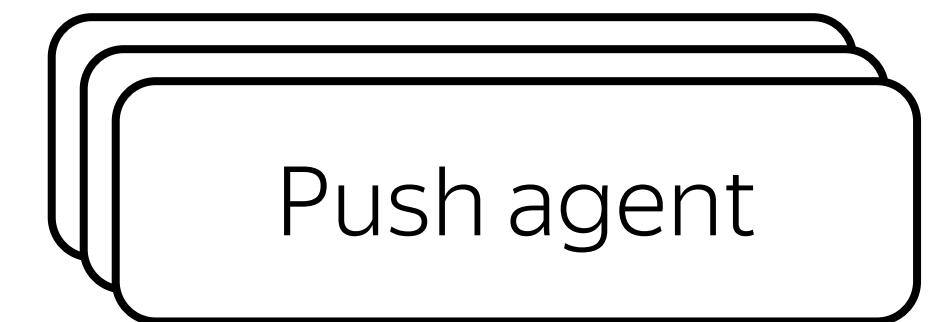
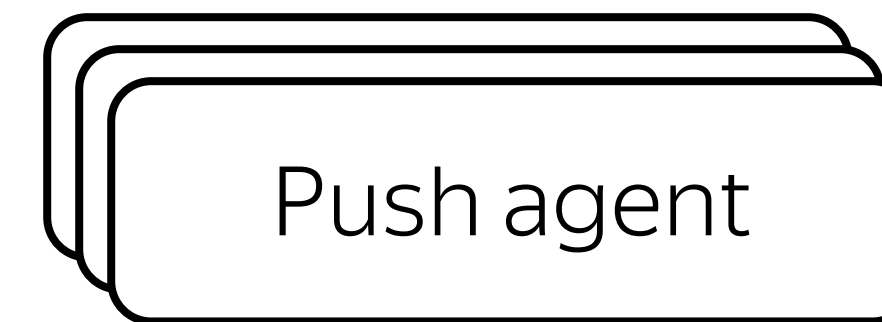
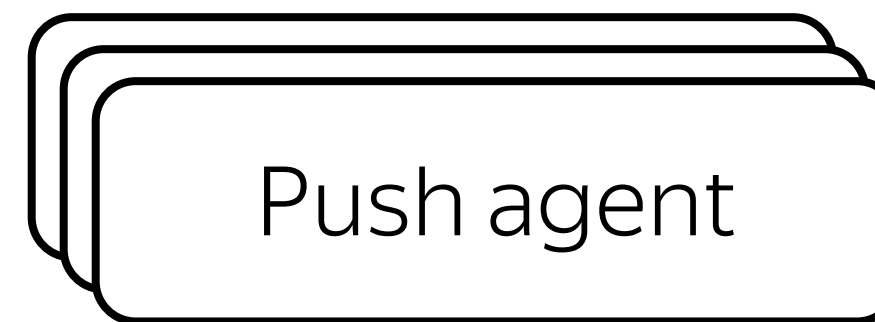
Поставка логов в DataLake



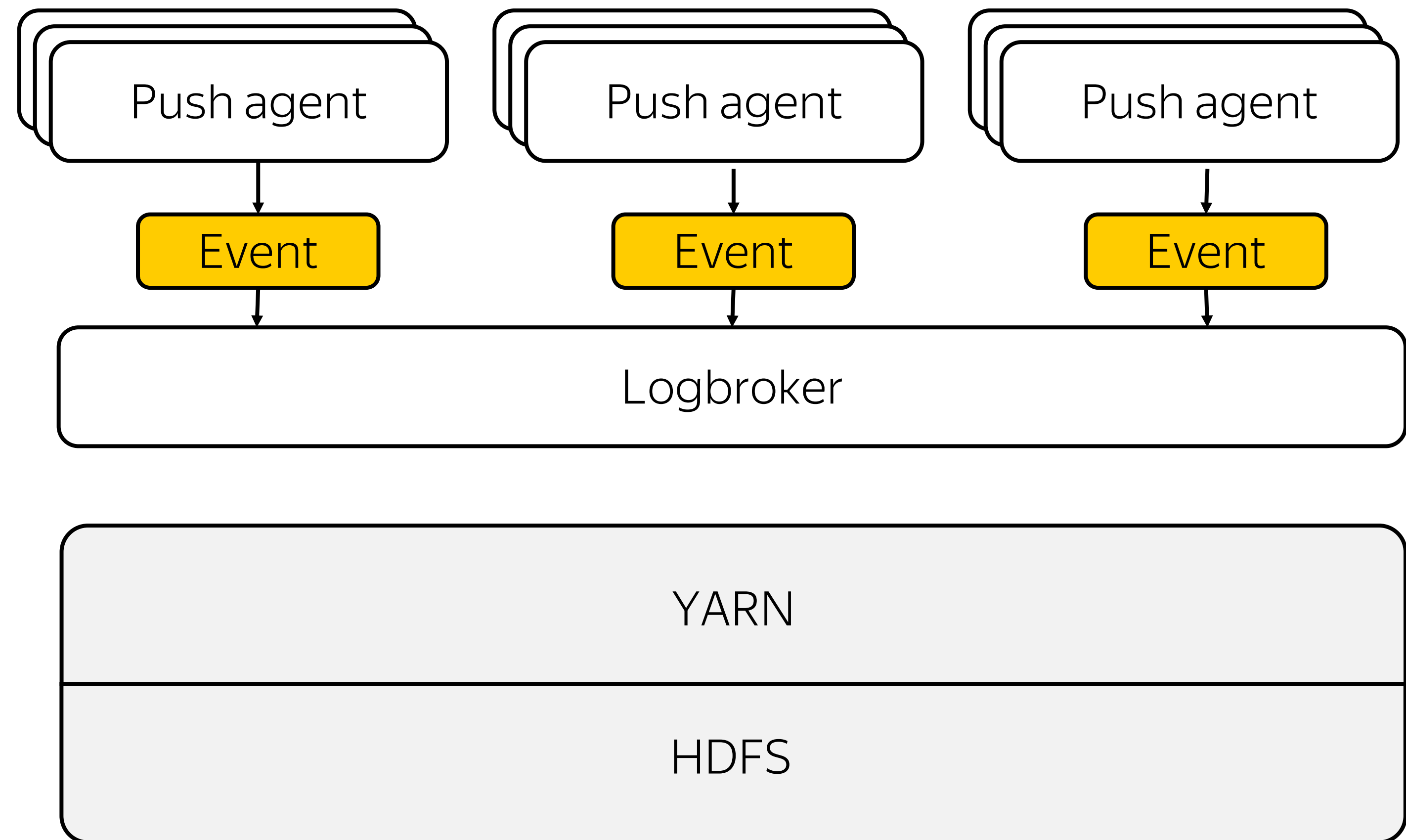
Поставка логов в DataLake



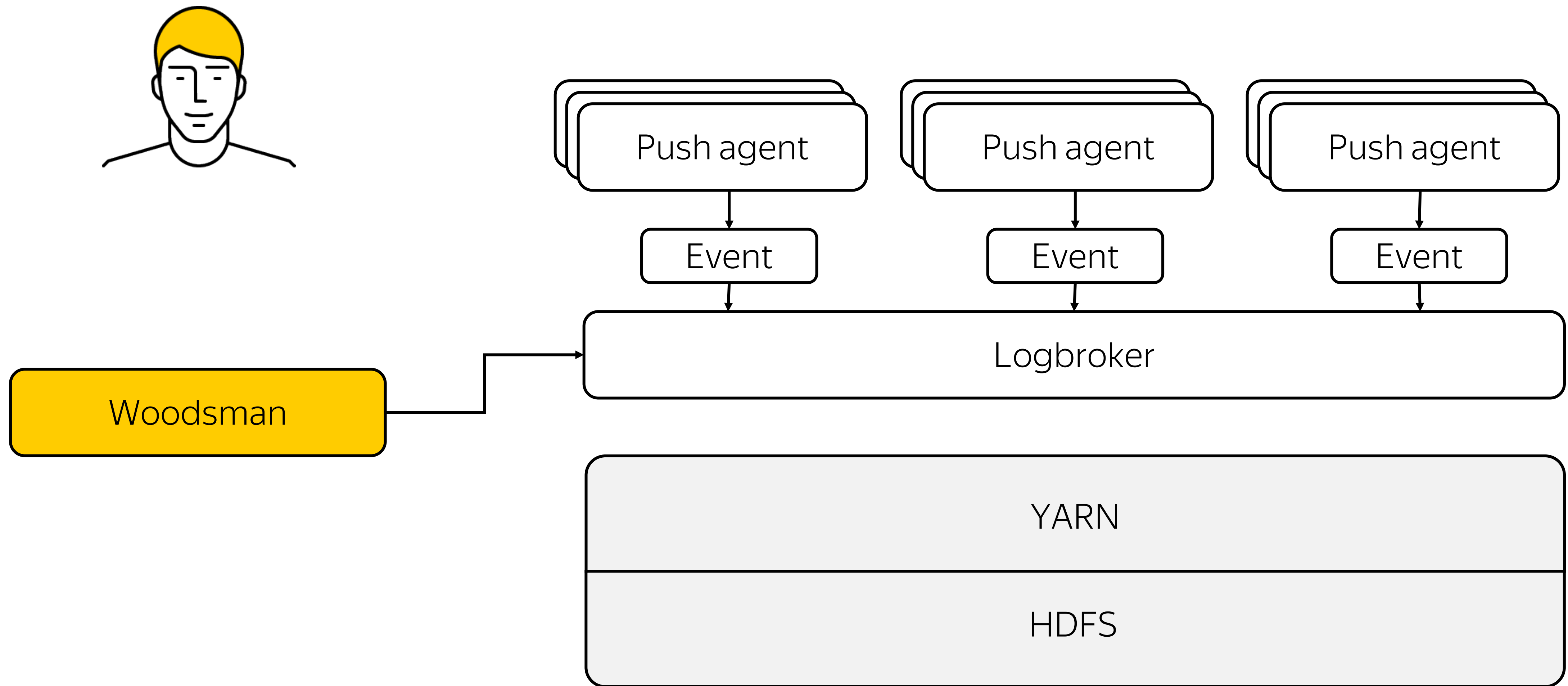
Поставка логов в DataLake



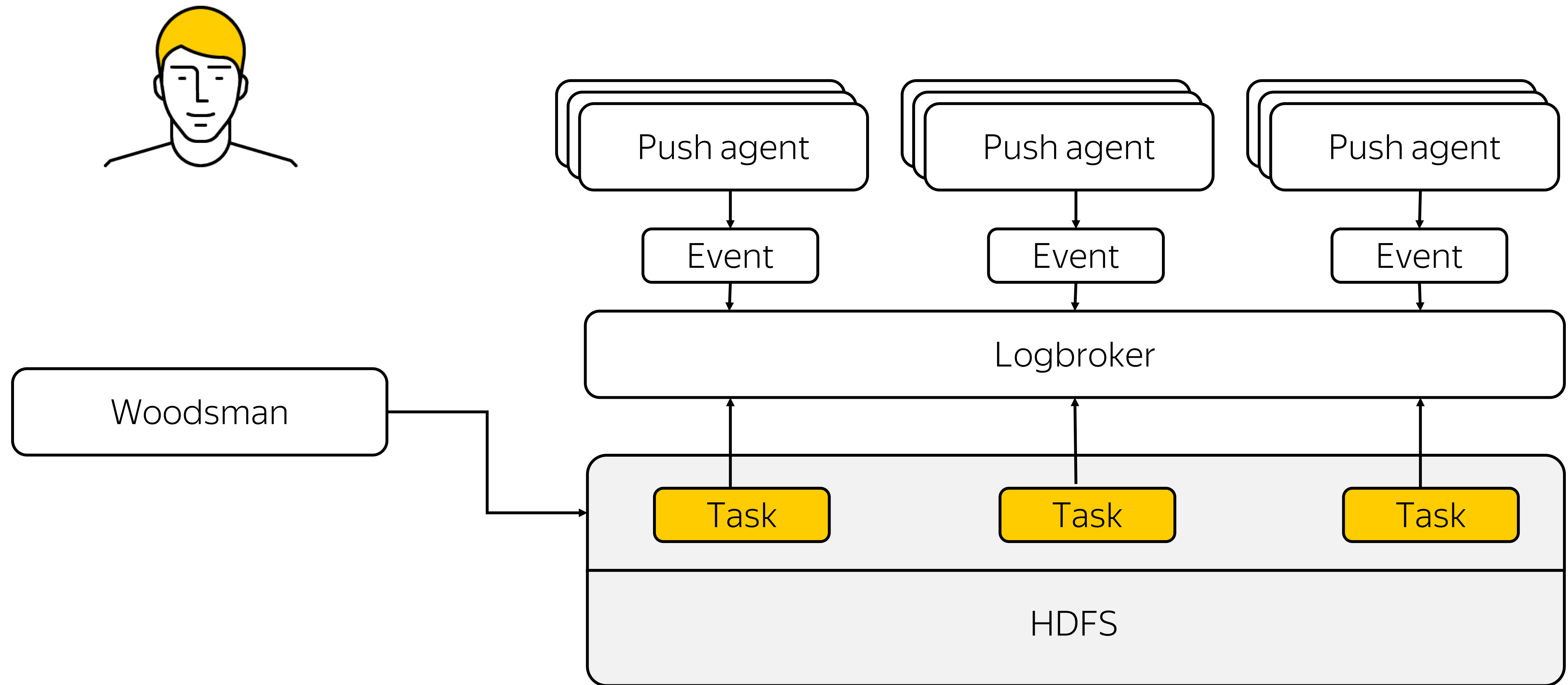
Поставка логов в DataLake



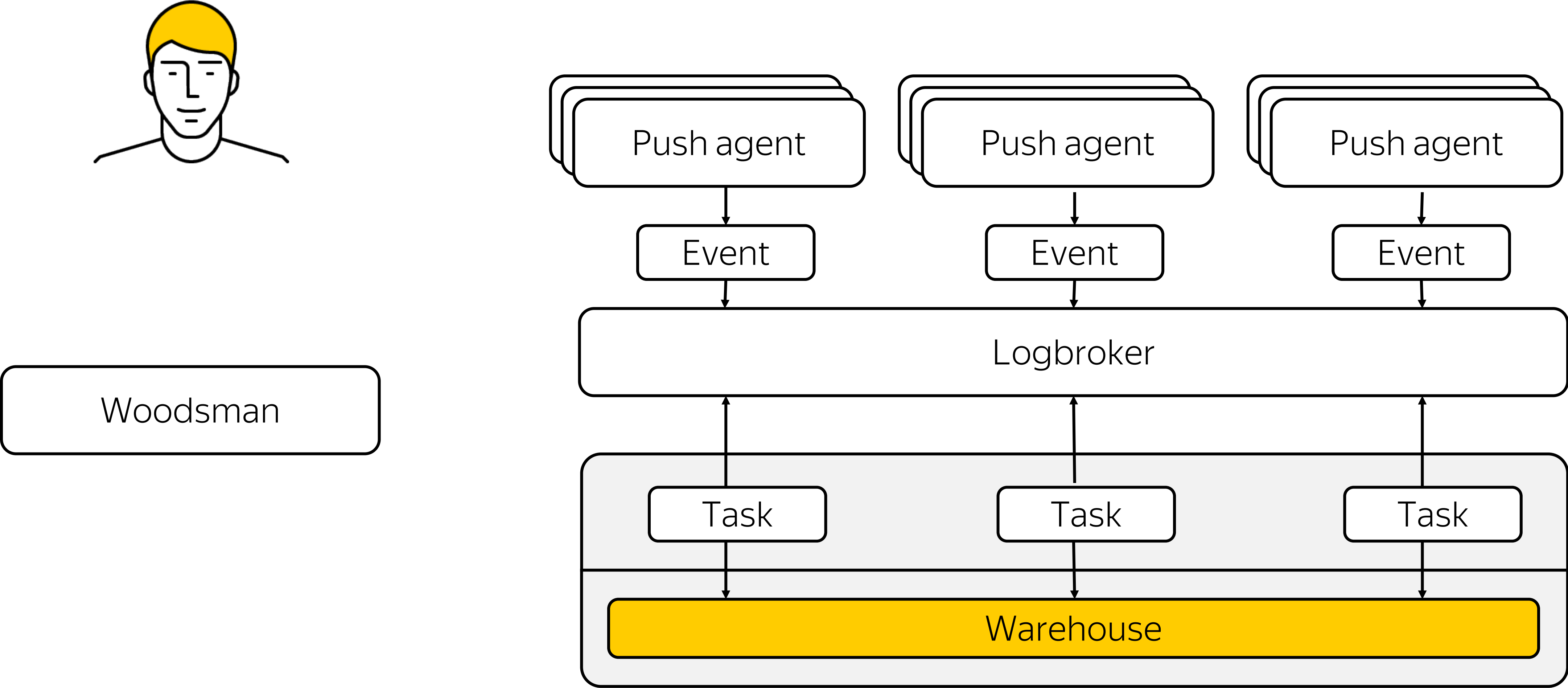
Поставка логов в DataLake



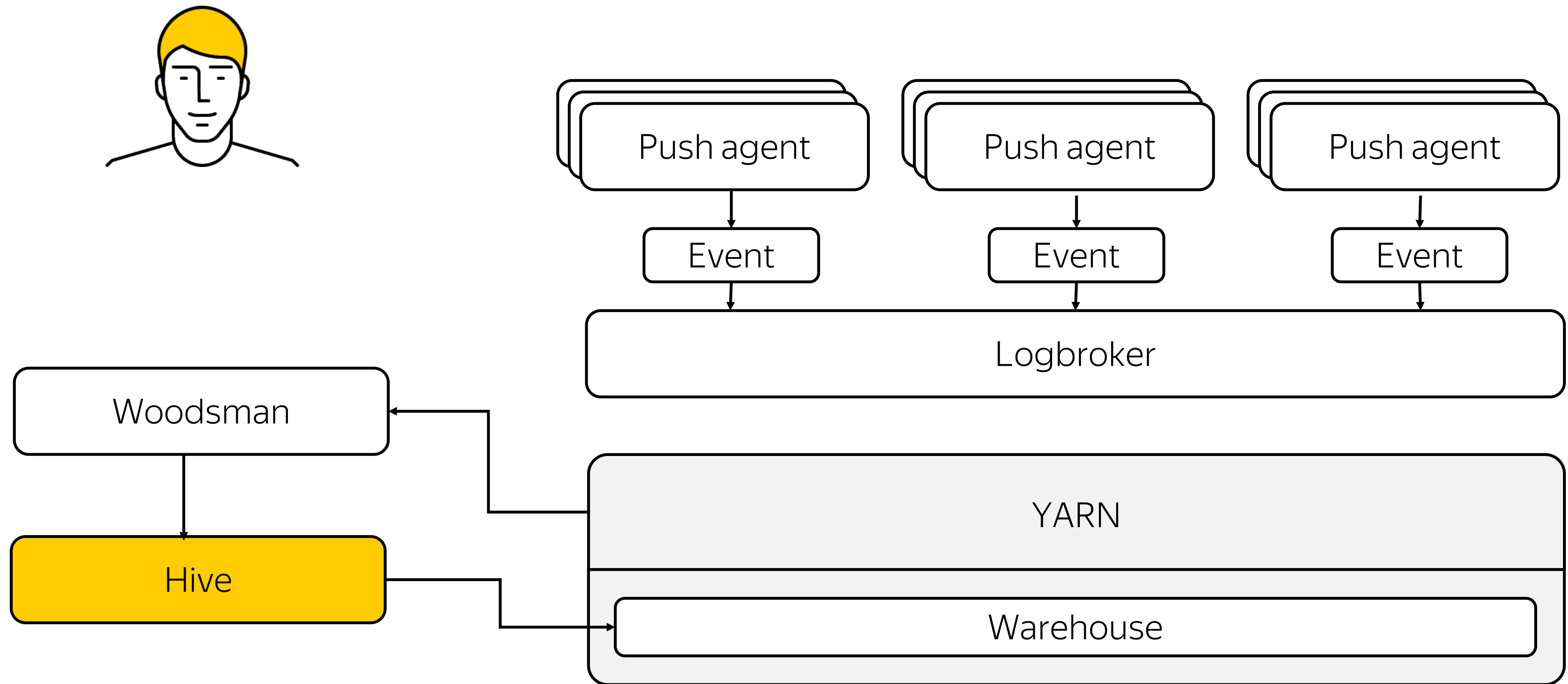
Поставка логов в DataLake



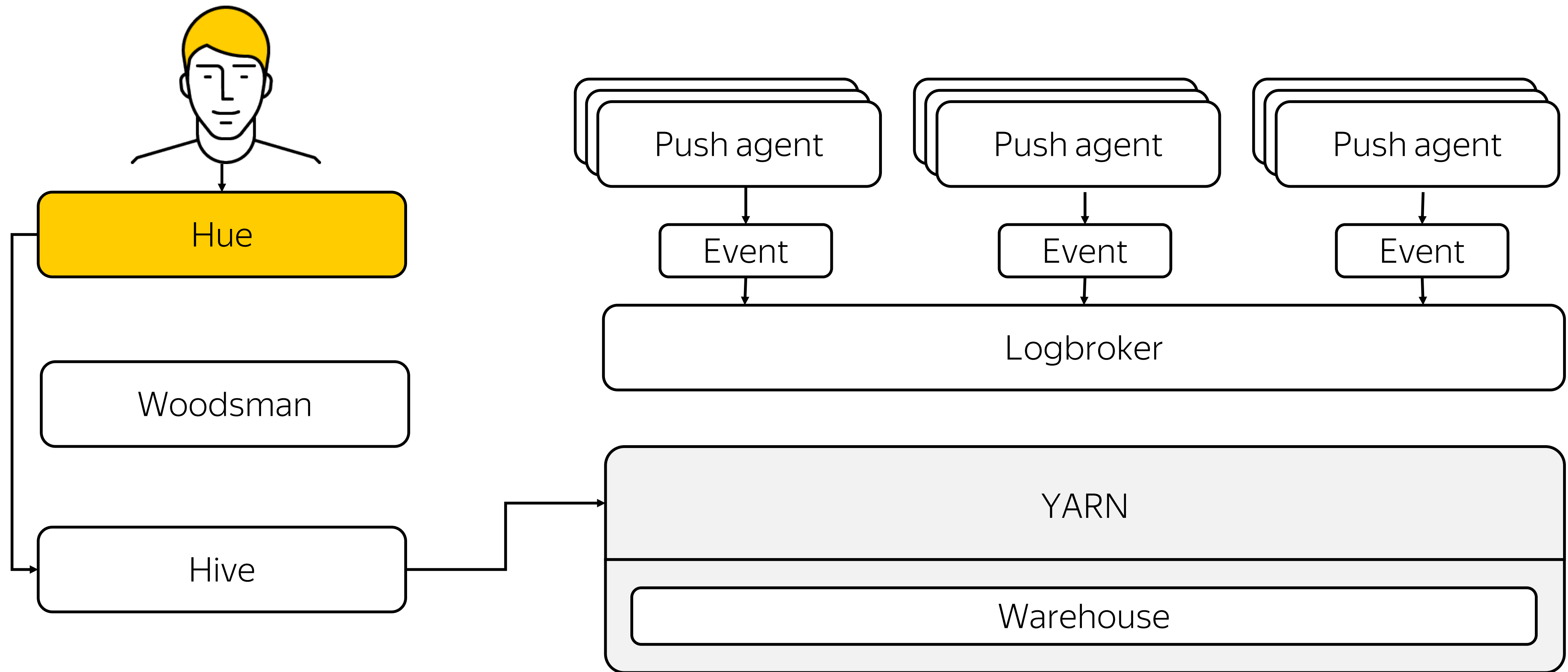
Поставка логов в DataLake



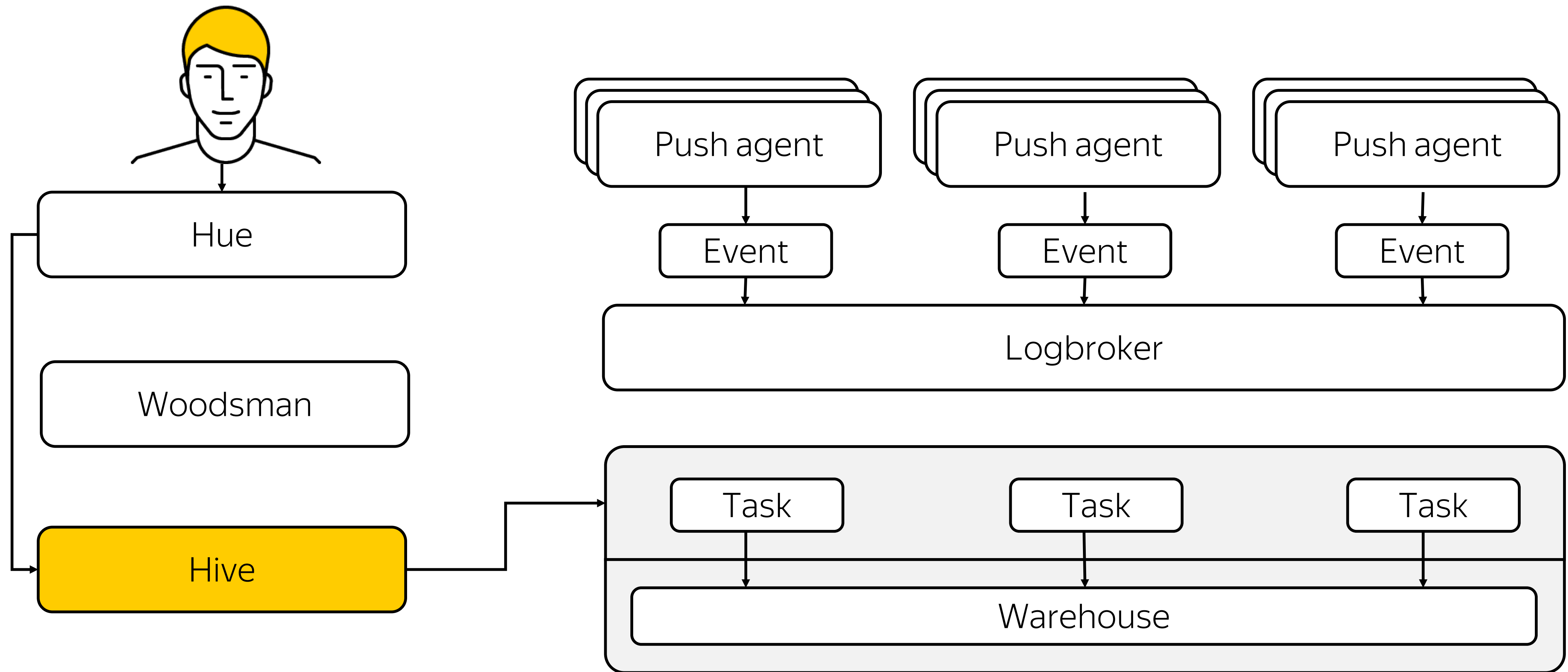
Поставка логов в DataLake



Поставка логов в DataLake



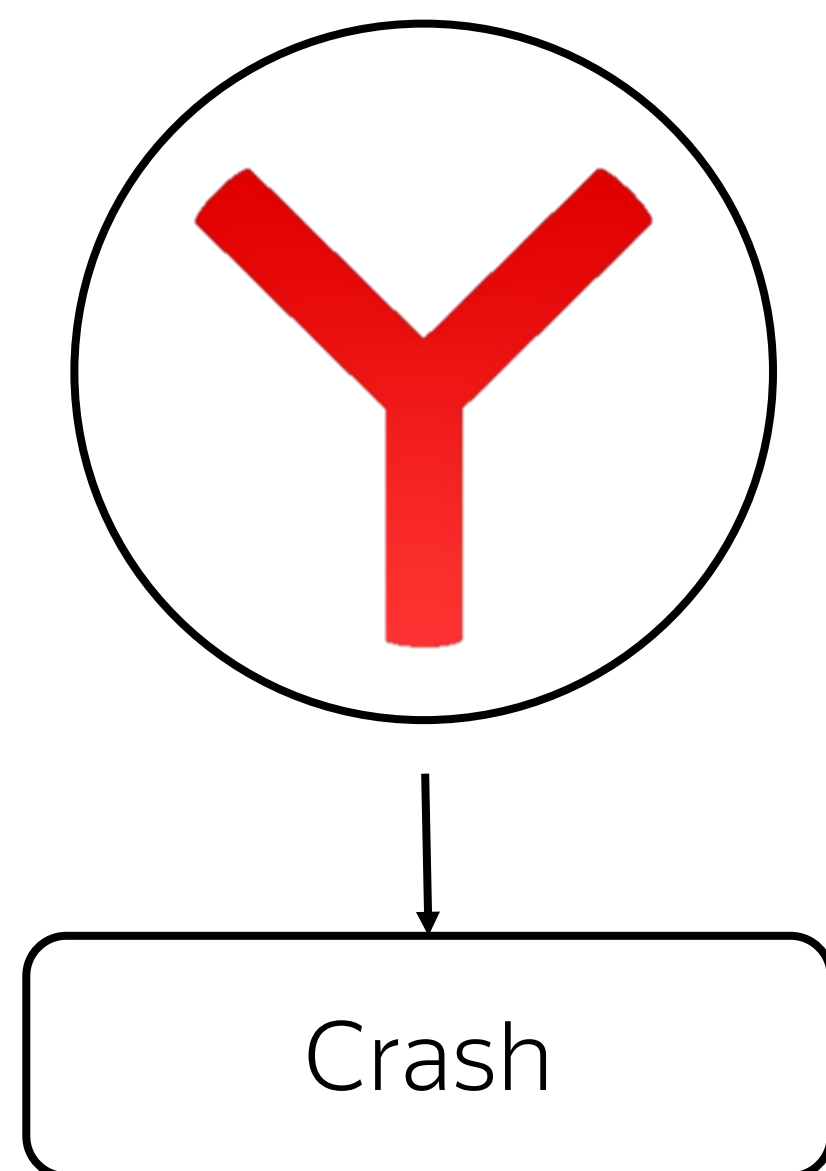
Поставка логов в DataLake



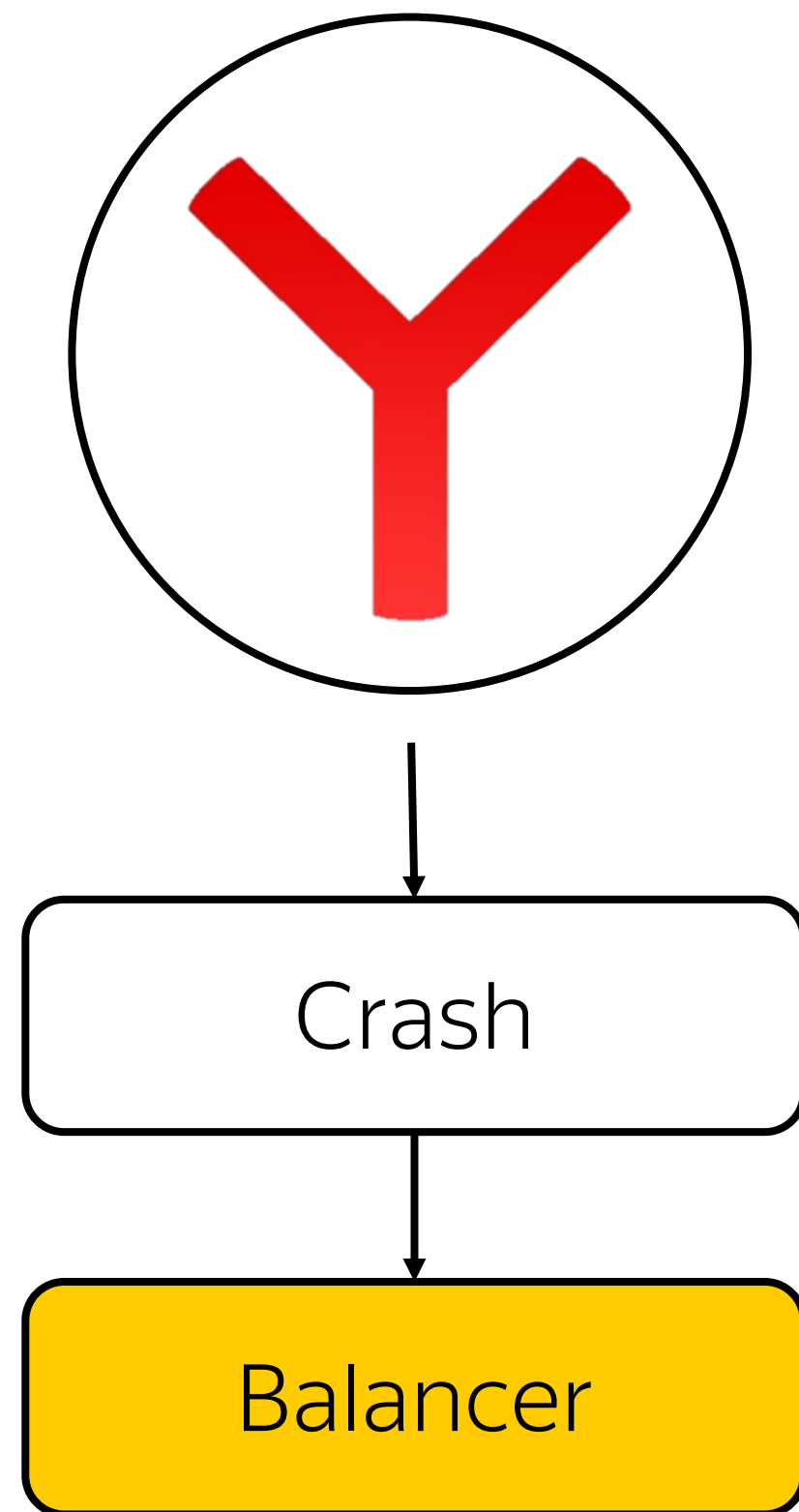
Яндекс.Браузер – система обработки крешей



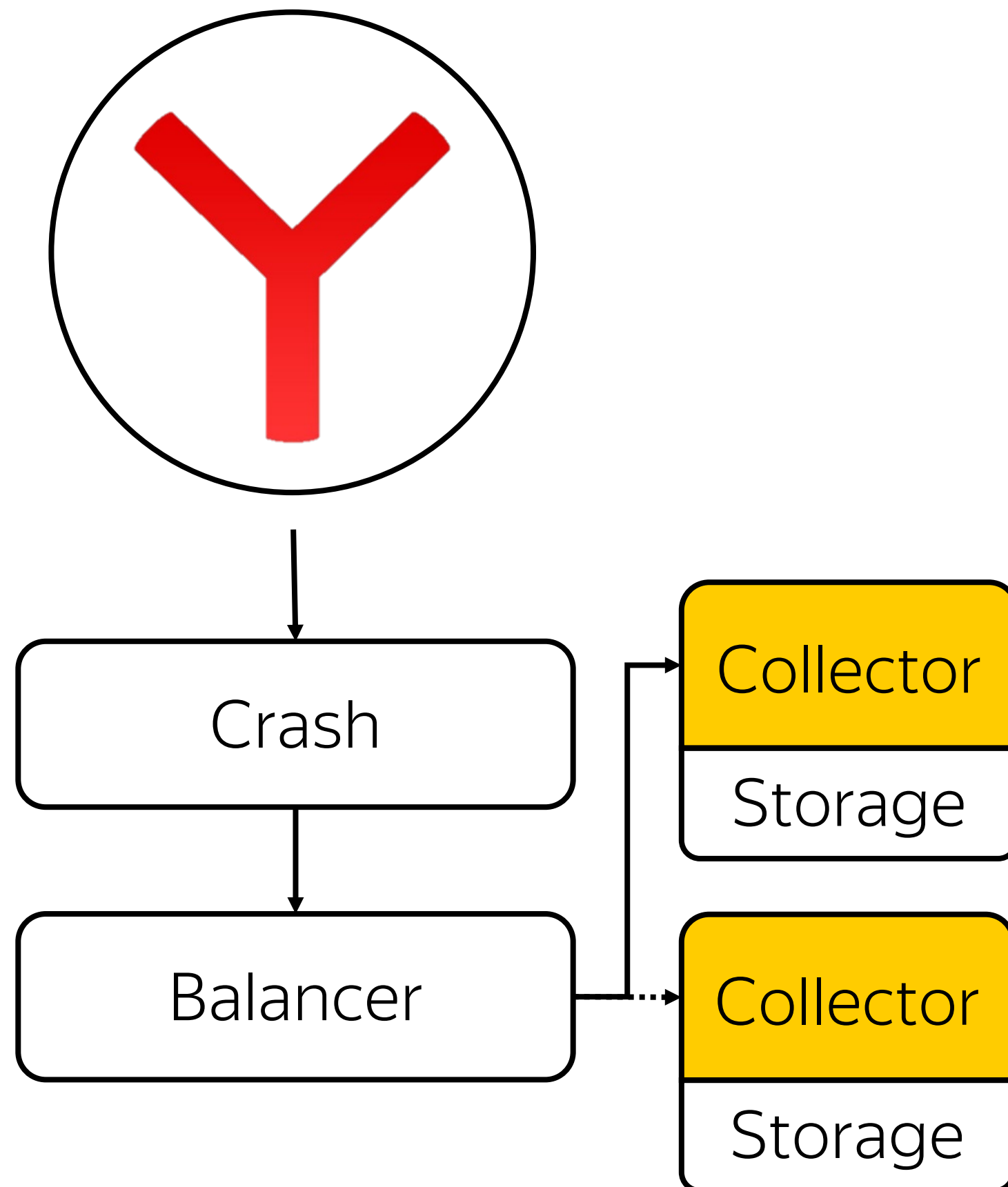
Яндекс.Браузер – система обработки крешей



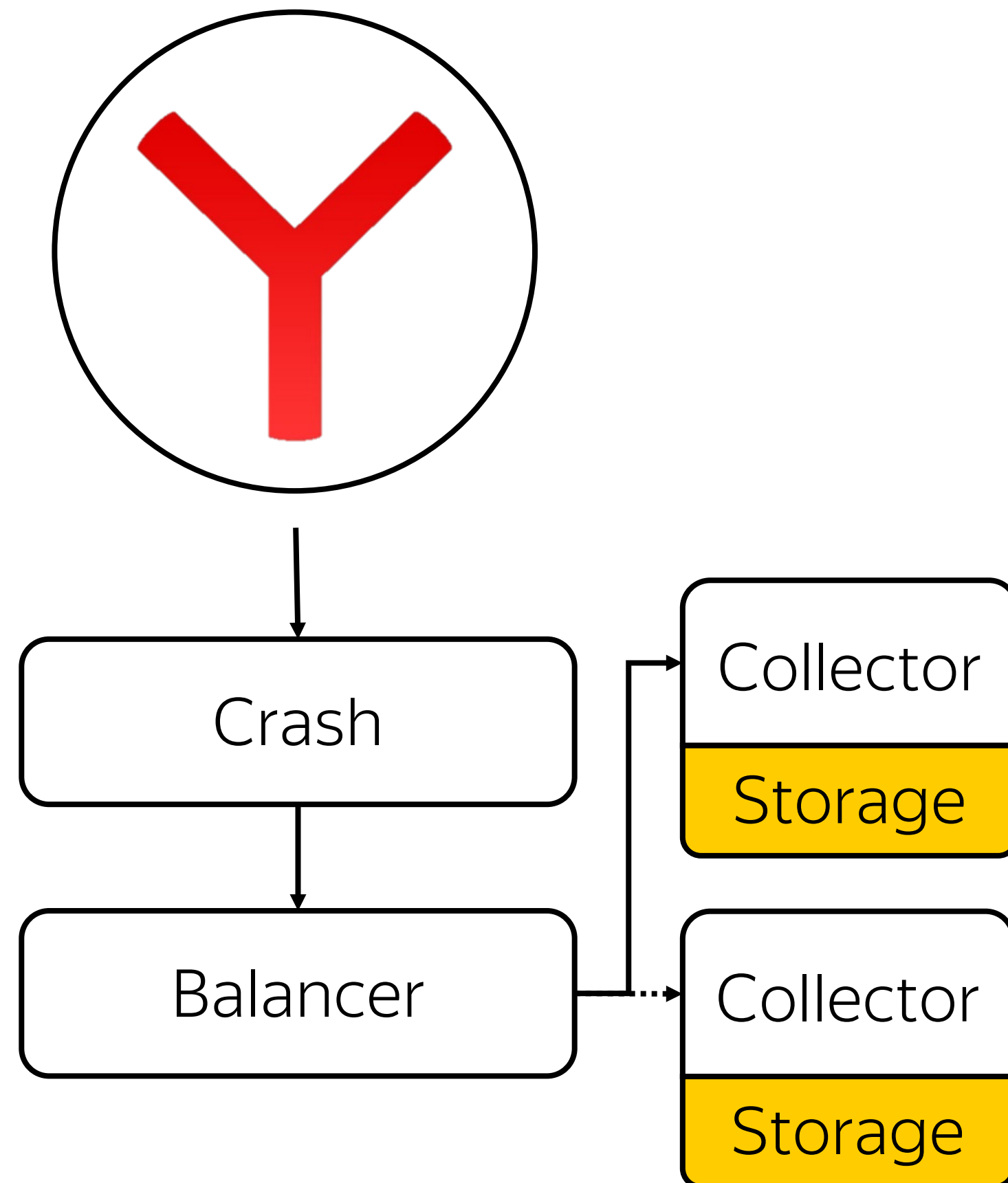
Яндекс.Браузер – система обработки крешей



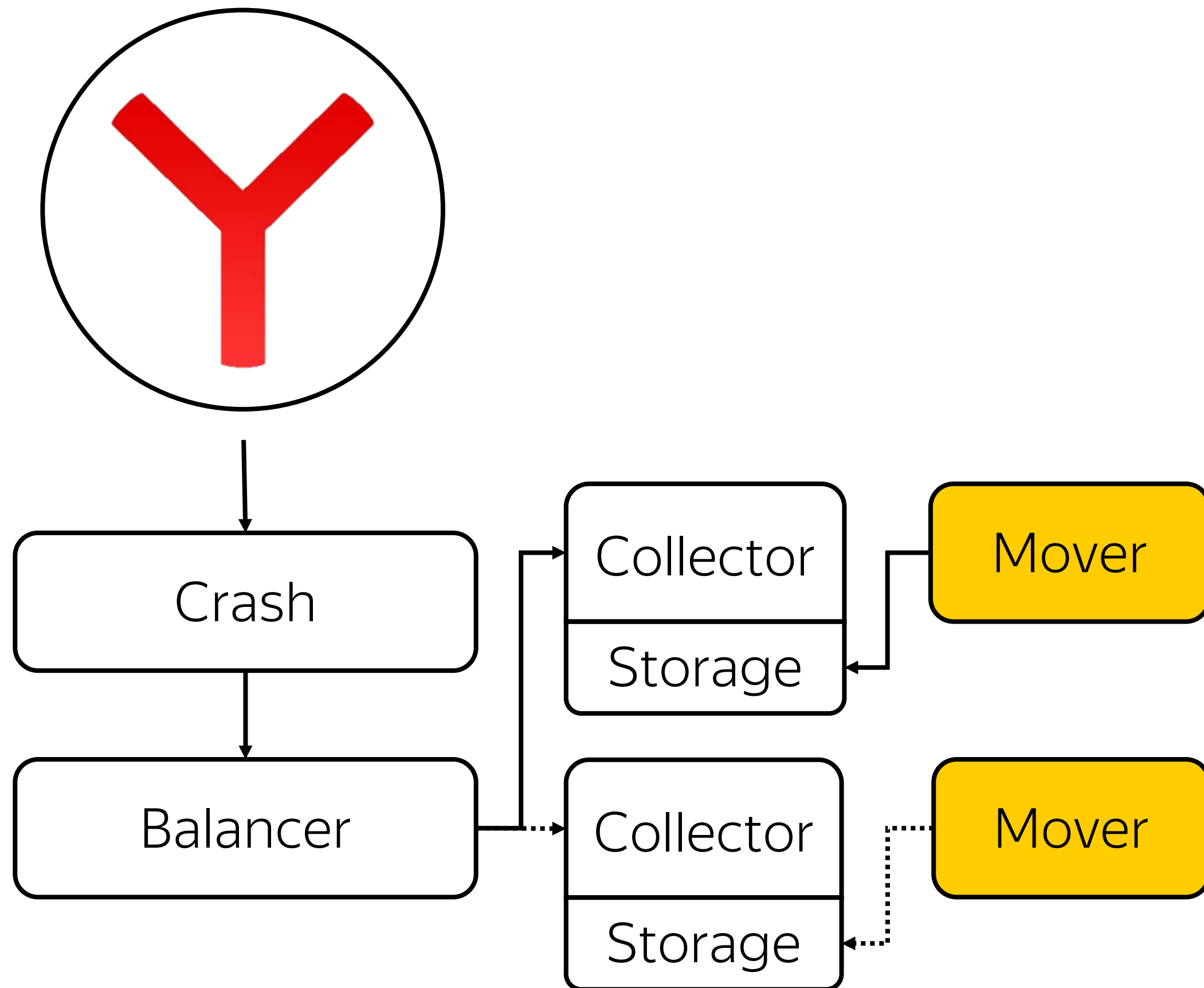
Яндекс.Браузер – система обработки крешей



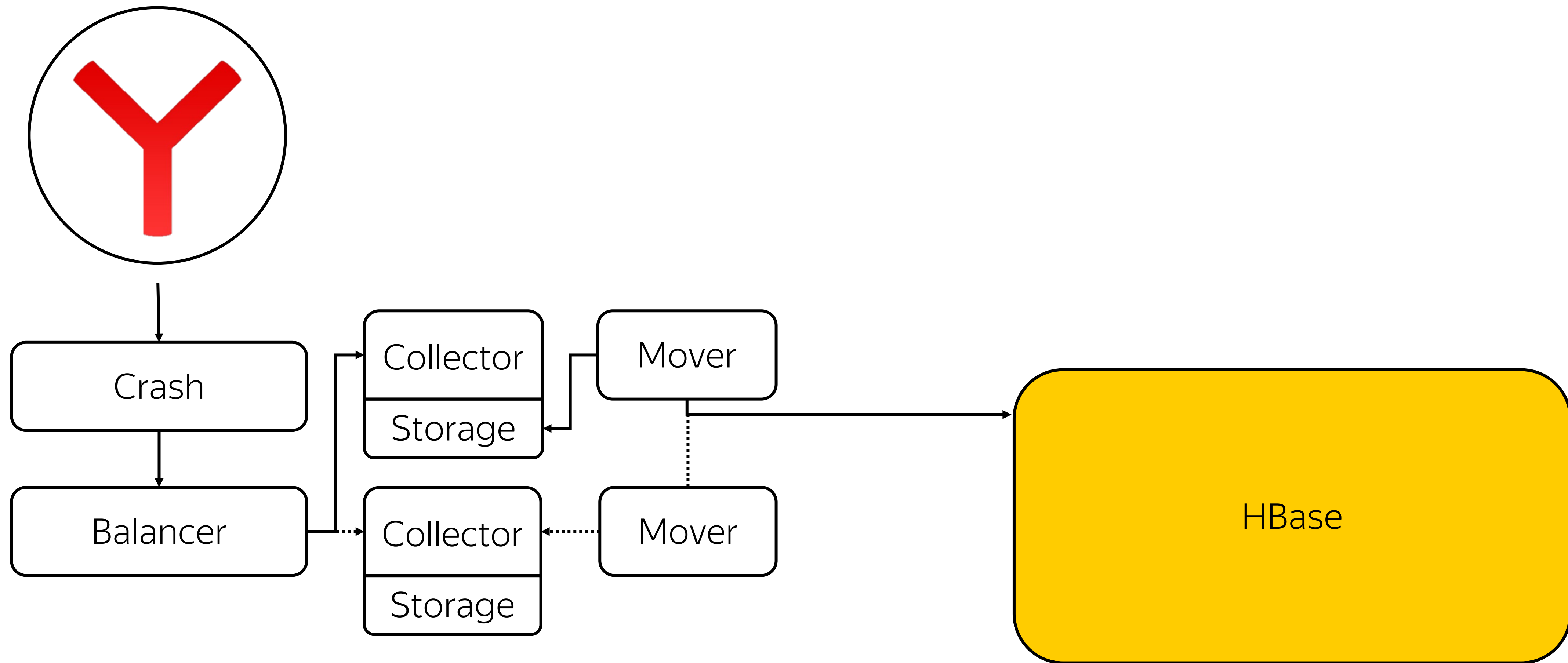
Яндекс.Браузер – система обработки крэшей



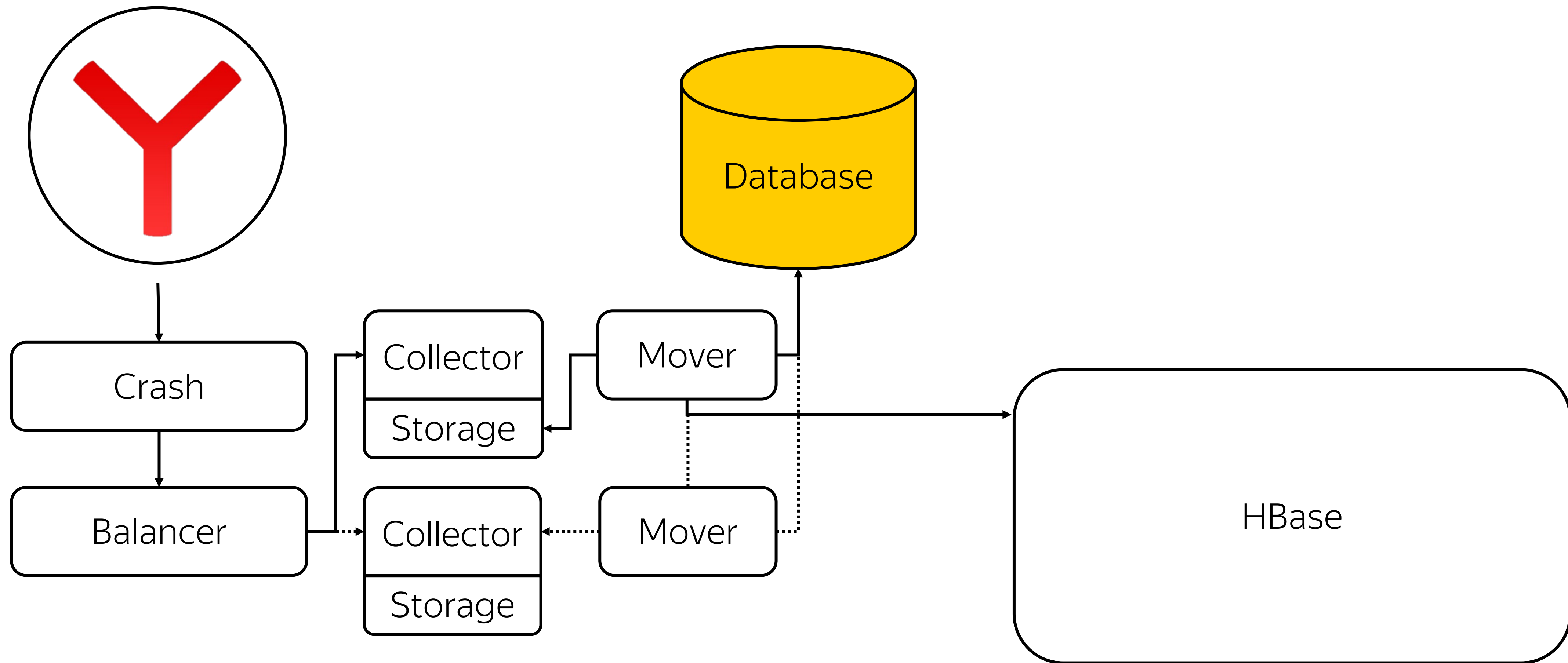
Яндекс.Браузер – система обработки крешей



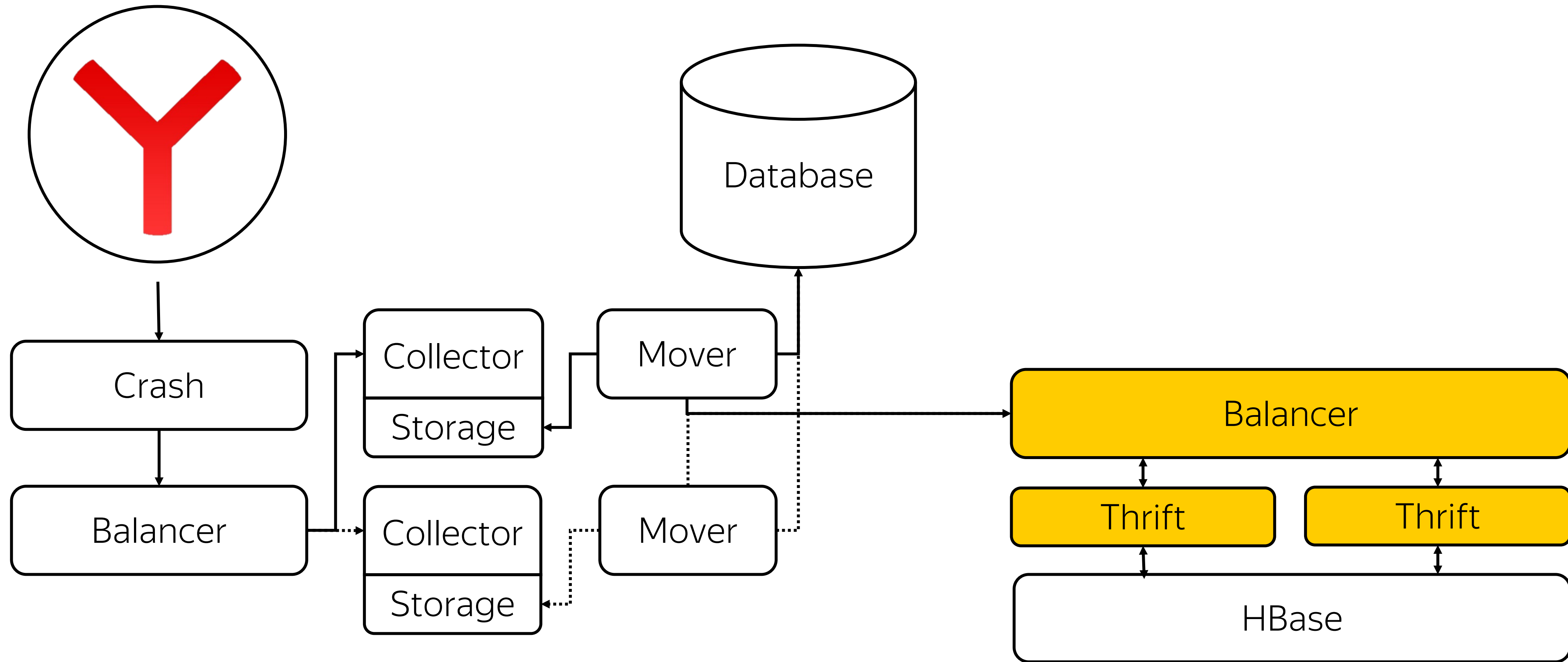
Яндекс.Браузер – система обработки крешей



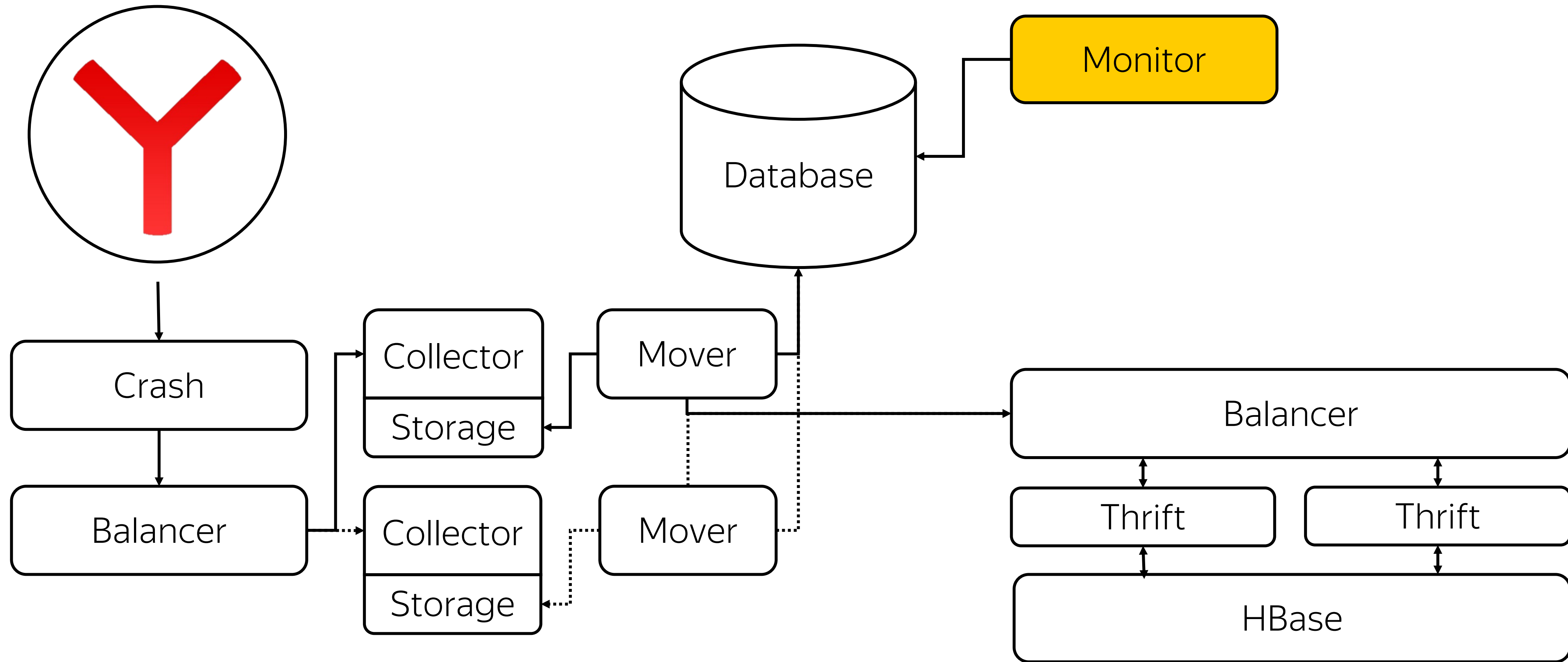
Яндекс.Браузер – система обработки крешей



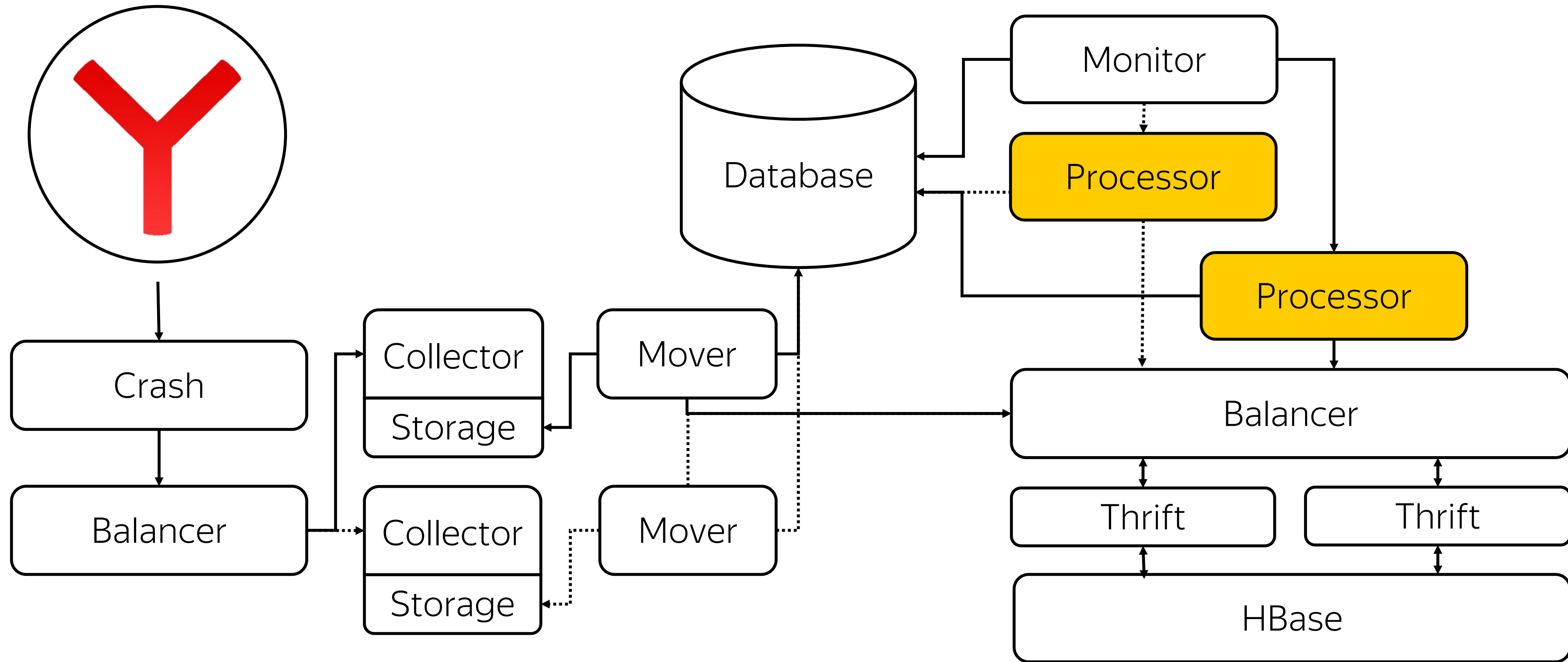
Яндекс.Браузер – система обработки крешей



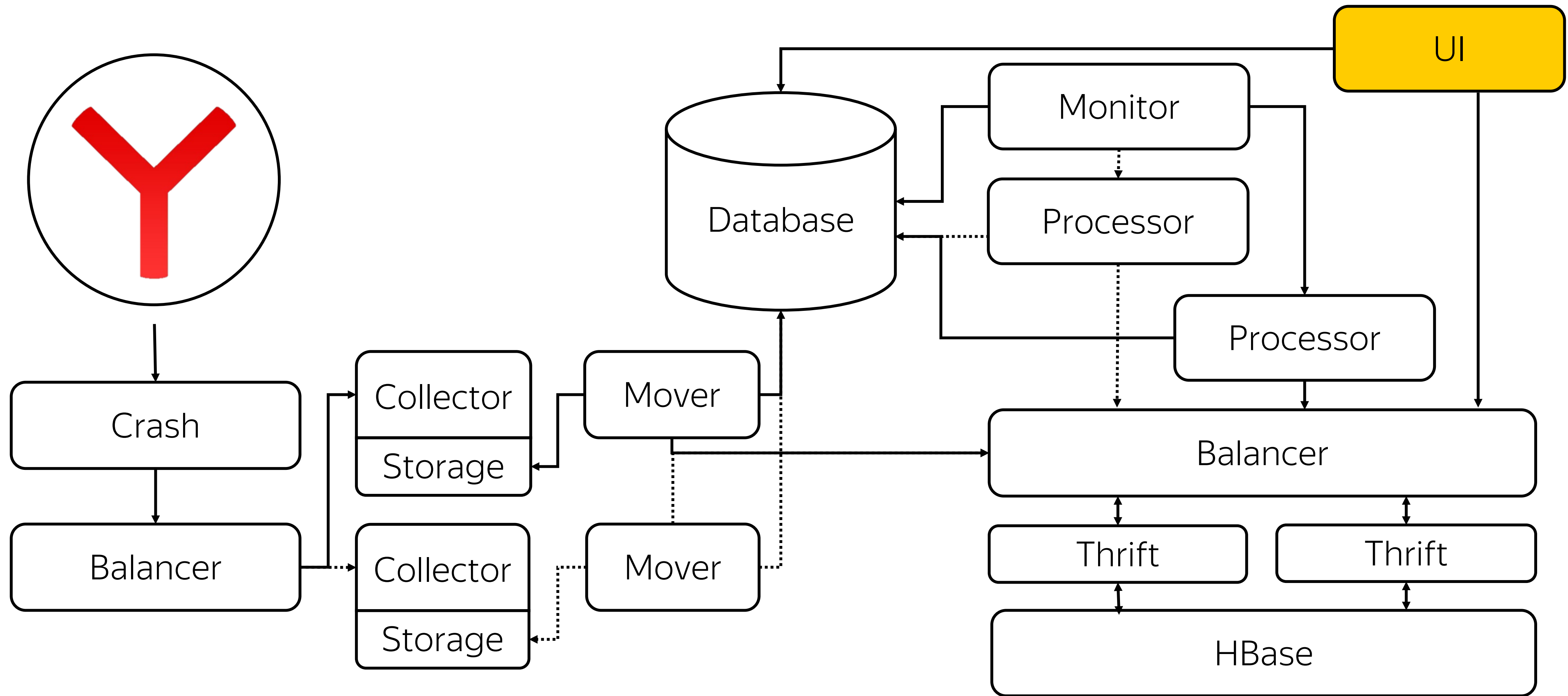
Яндекс.Браузер – система обработки крашей



Яндекс.Браузер – система обработки крешей



Яндекс.Браузер – система обработки крешей



Яндекс.Маркет – поставка и анализ логов

Подробный рассказ

<https://events.yandex.ru/lib/talks/3713>

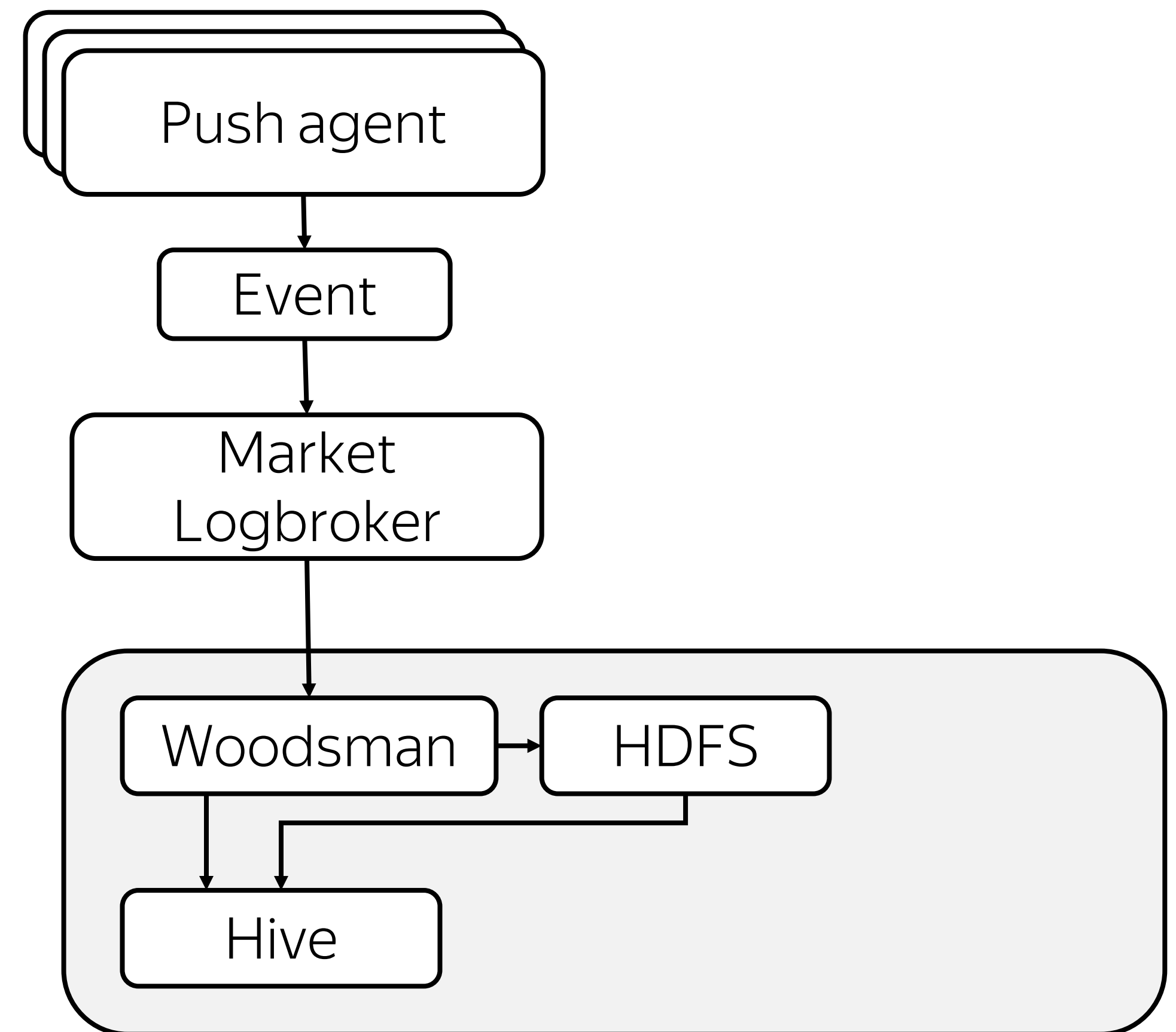
- › Поток событий в Hive
- › ETL и Ad-hoc запросы аналитиков

Яндекс.Маркет – поставка и анализ логов

Подробный рассказ

<https://events.yandex.ru/lib/talks/3713>

- › Поток событий в Hive
- › ETL и Ad-hoc запросы аналитиков

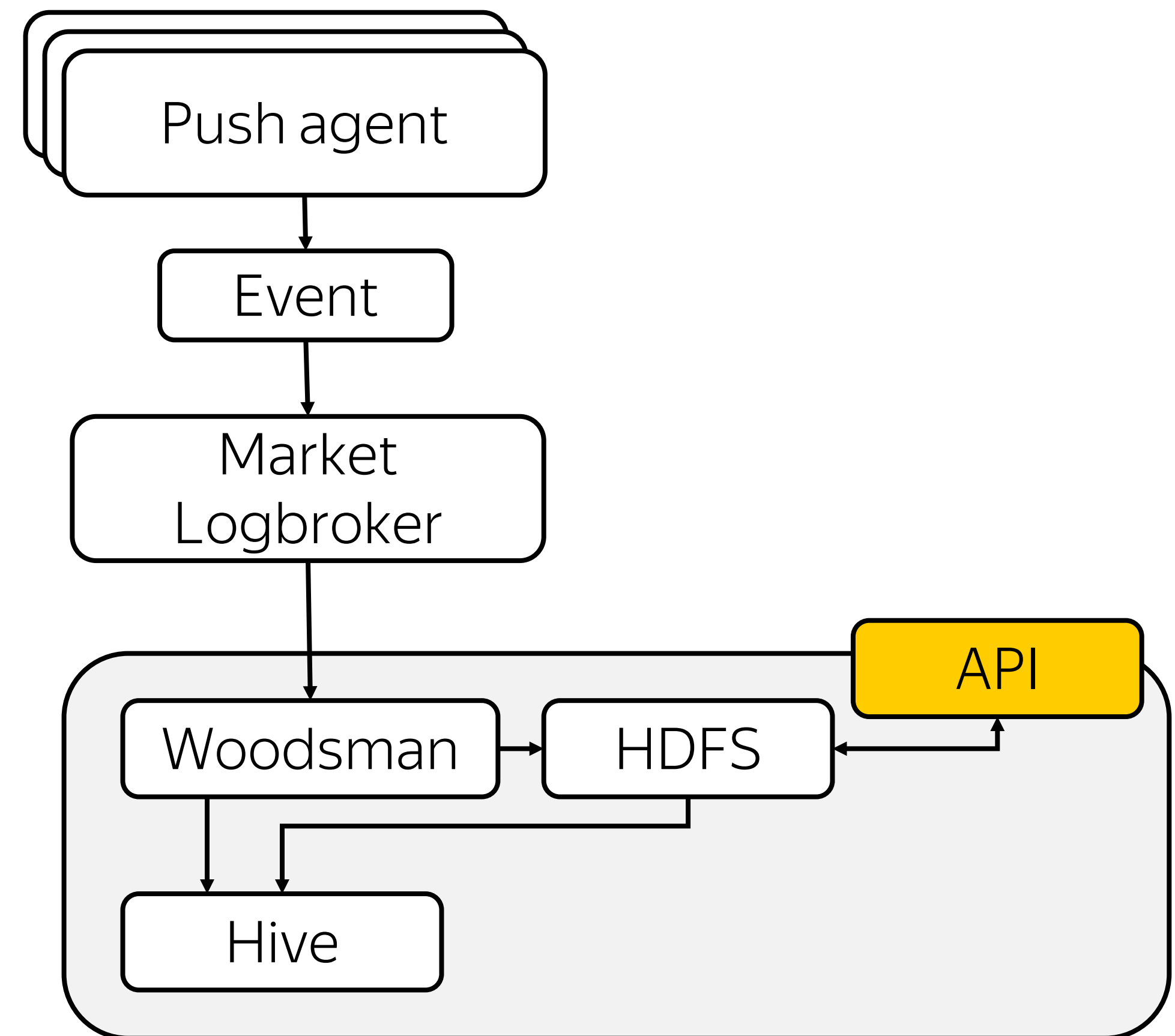


Яндекс.Маркет – поставка и анализ логов

Подробный рассказ

<https://events.yandex.ru/lib/talks/3713>

- › Поток событий в Hive
- › ETL и Ad-hoc запросы аналитиков

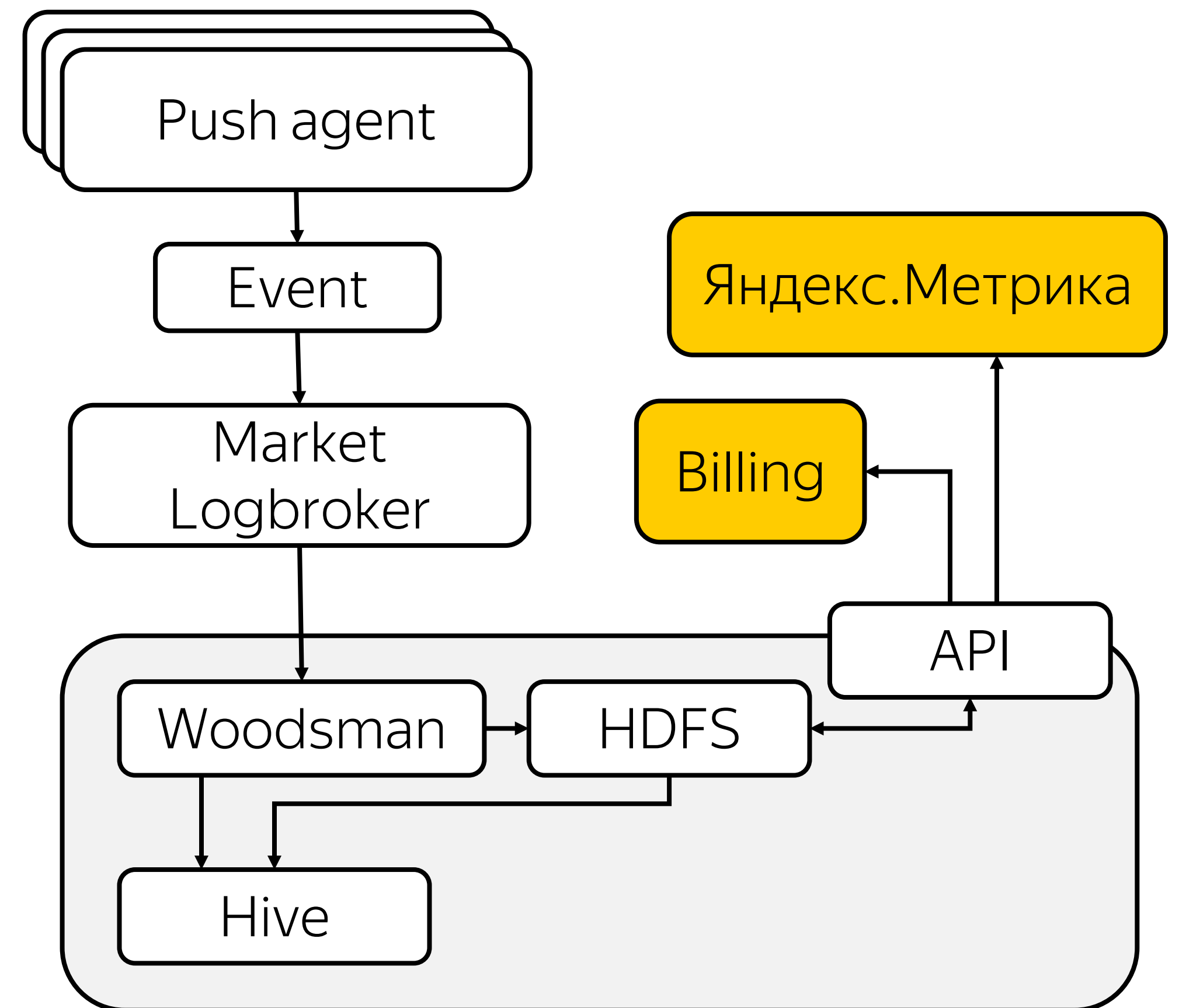


Яндекс.Маркет – поставка и анализ логов

Подробный рассказ

<https://events.yandex.ru/lib/talks/3713>

- › Поток событий в Hive
- › ETL и Ad-hoc запросы аналитиков

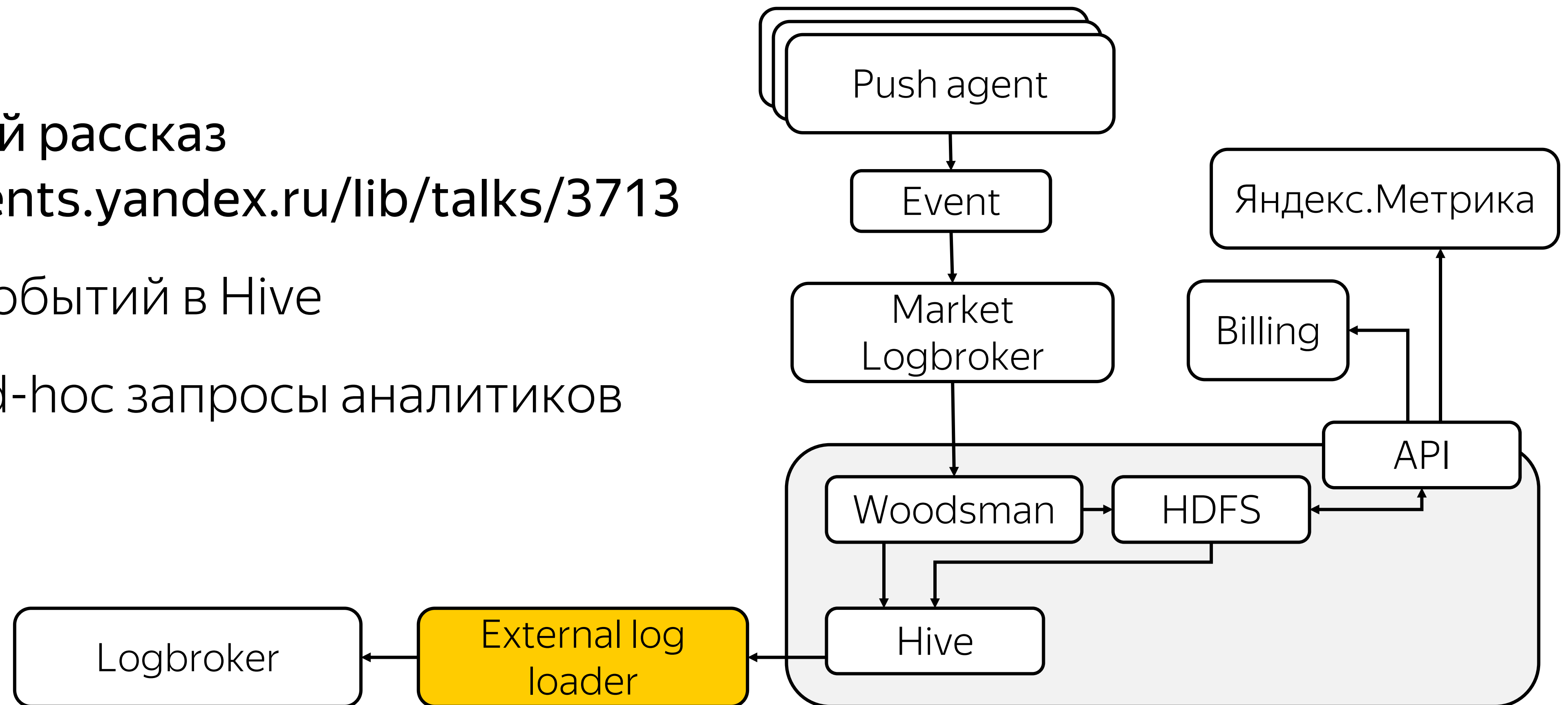


Яндекс.Маркет – поставка и анализ логов

Подробный рассказ

<https://events.yandex.ru/lib/talks/3713>

- › Поток событий в Hive
- › ETL и Ad-hoc запросы аналитиков

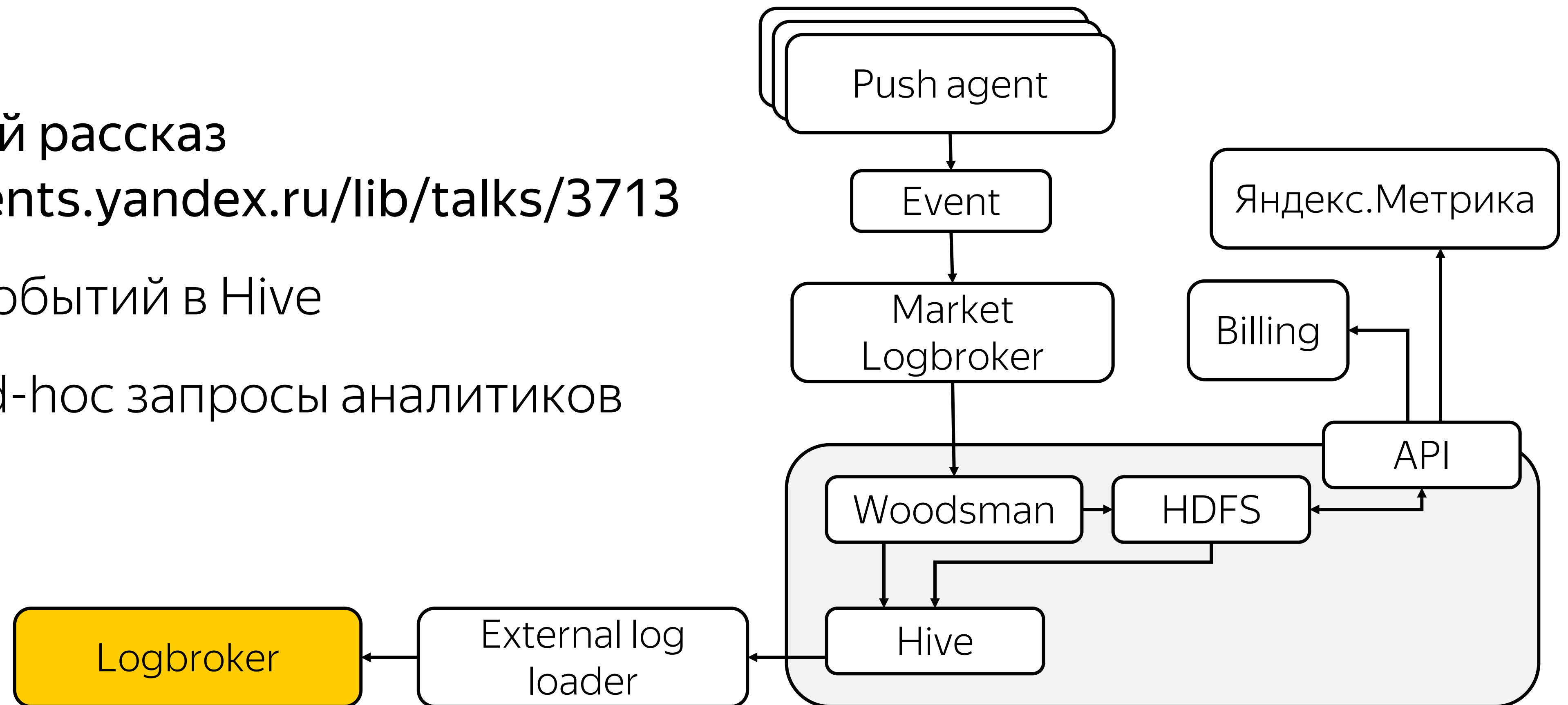


Яндекс.Маркет – поставка и анализ логов

Подробный рассказ

<https://events.yandex.ru/lib/talks/3713>

- › Поток событий в Hive
- › ETL и Ad-hoc запросы аналитиков

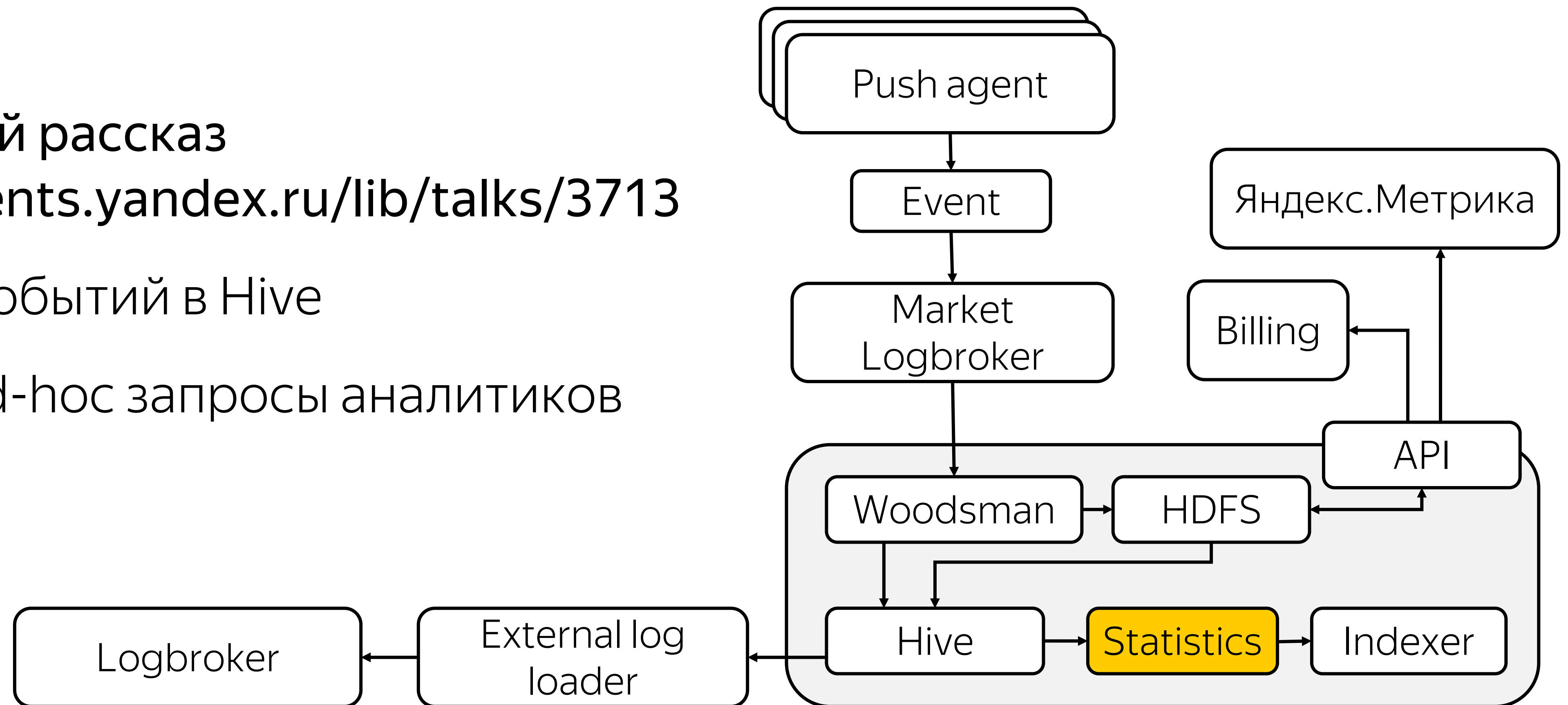


Яндекс.Маркет – поставка и анализ логов

Подробный рассказ

<https://events.yandex.ru/lib/talks/3713>

- › Поток событий в Hive
- › ETL и Ad-hoc запросы аналитиков

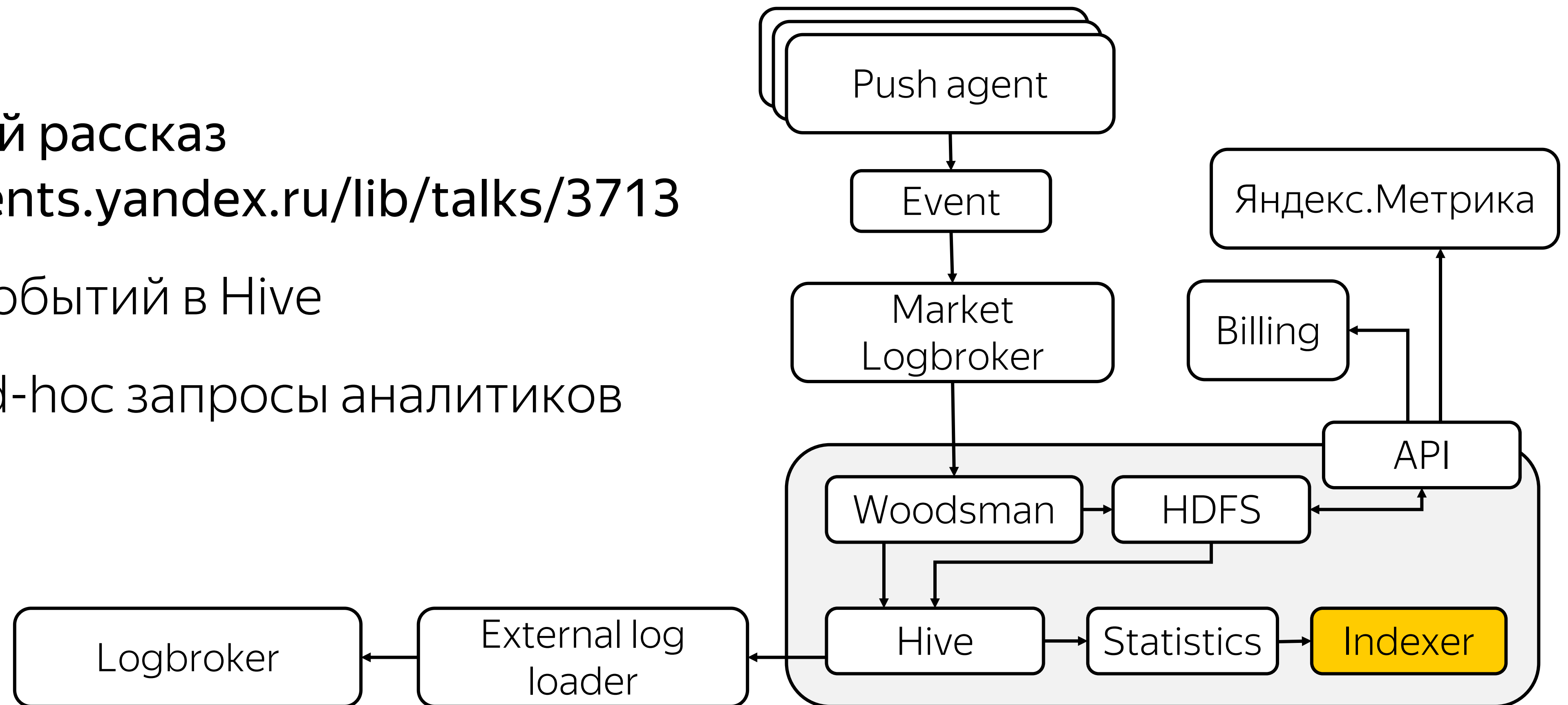


Яндекс.Маркет – поставка и анализ логов

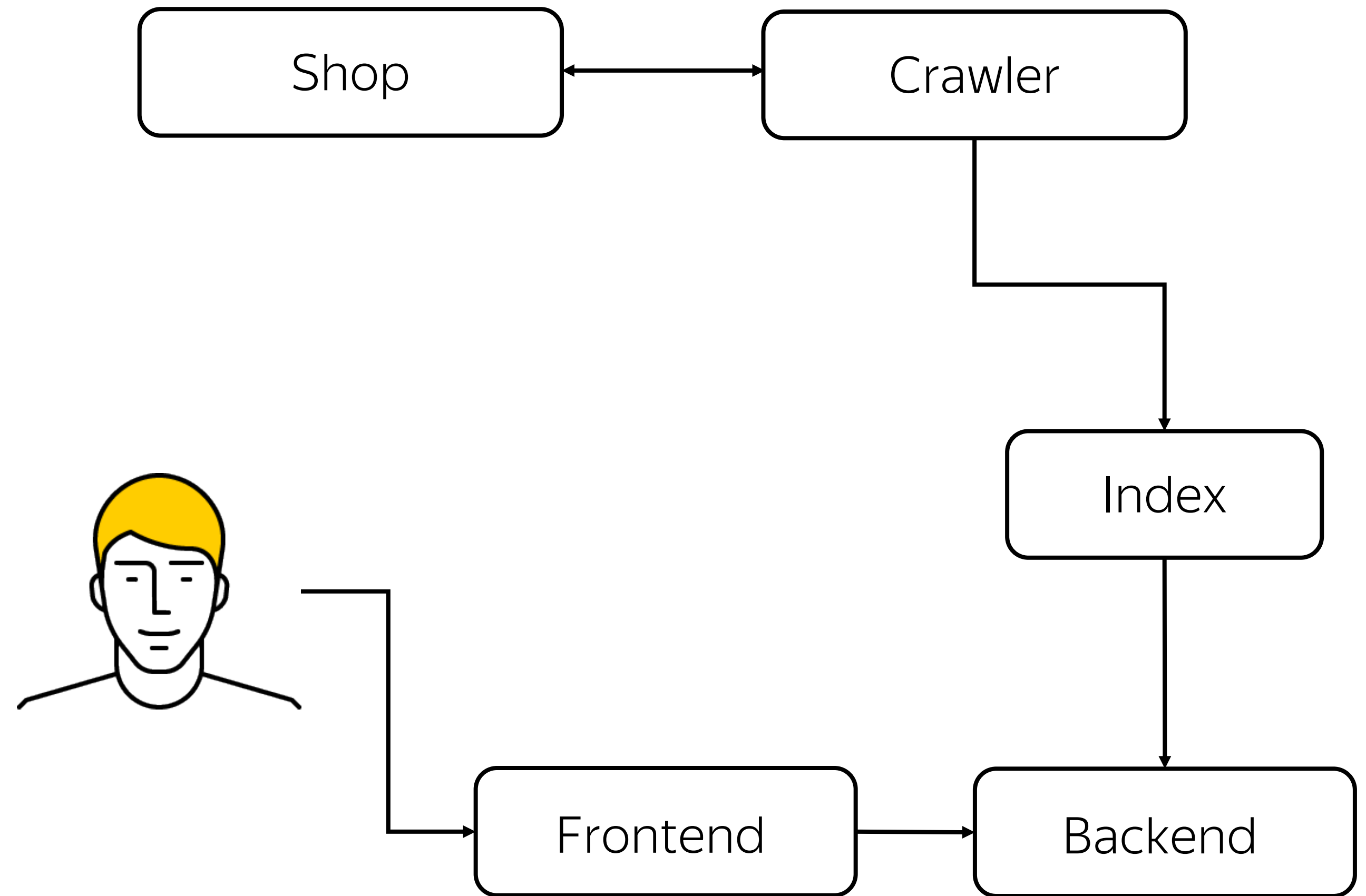
Подробный рассказ

<https://events.yandex.ru/lib/talks/3713>

- › Поток событий в Hive
- › ETL и Ad-hoc запросы аналитиков



Яндекс.Маркет – построение индекса



Яндекс.Маркет – построение индекса

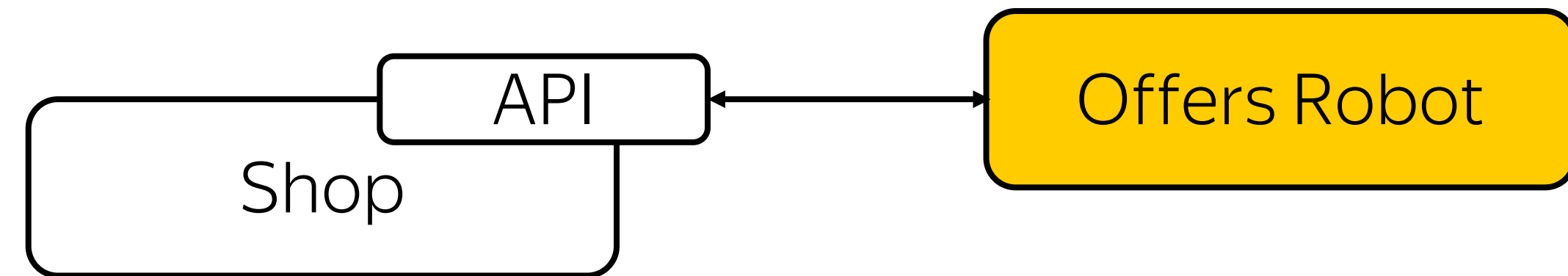
Магазины экспортируют
список предложений в API



Яндекс.Маркет – построение индекса

Магазины экспортируют
список предложений в API

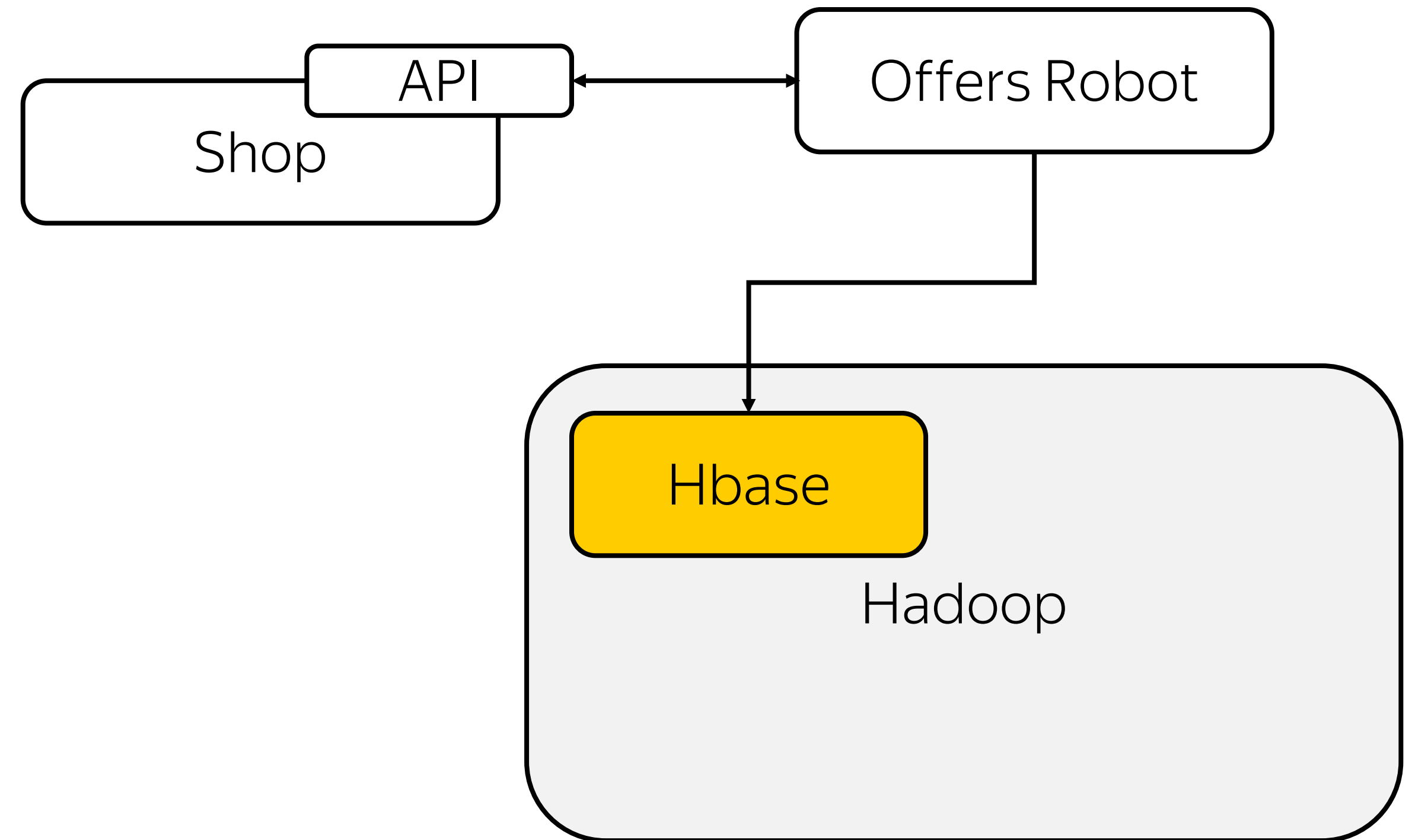
Offers robot собирает
предложения в hbase



Яндекс.Маркет – построение индекса

Магазины экспортируют список предложений в API

Offers robot собирает предложения в hbase

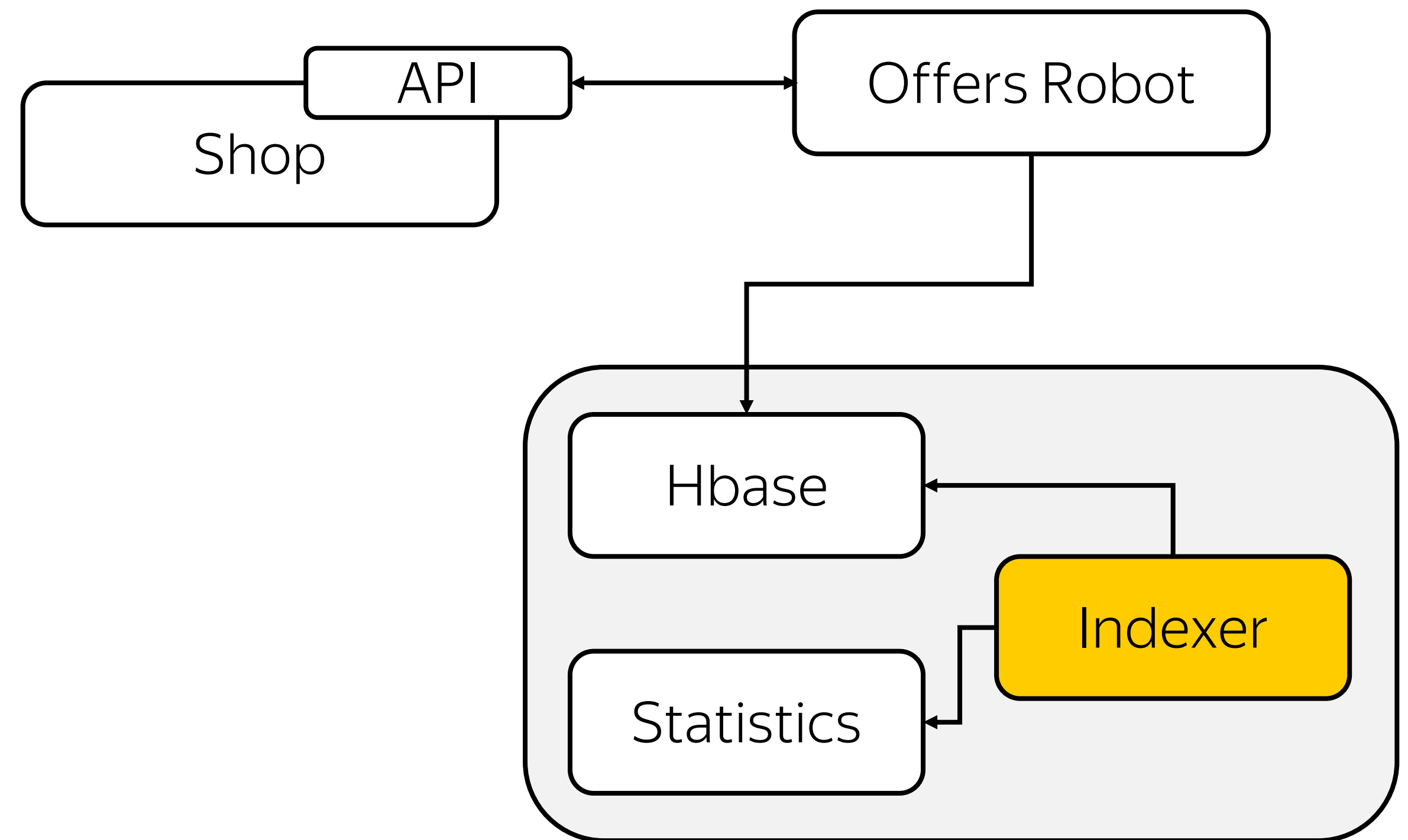


Яндекс.Маркет – построение индекса

Магазины экспортируют список предложений в API

Offers robot собирает предложения в hbase

Indexer вычитывает таблицу и строит индекс



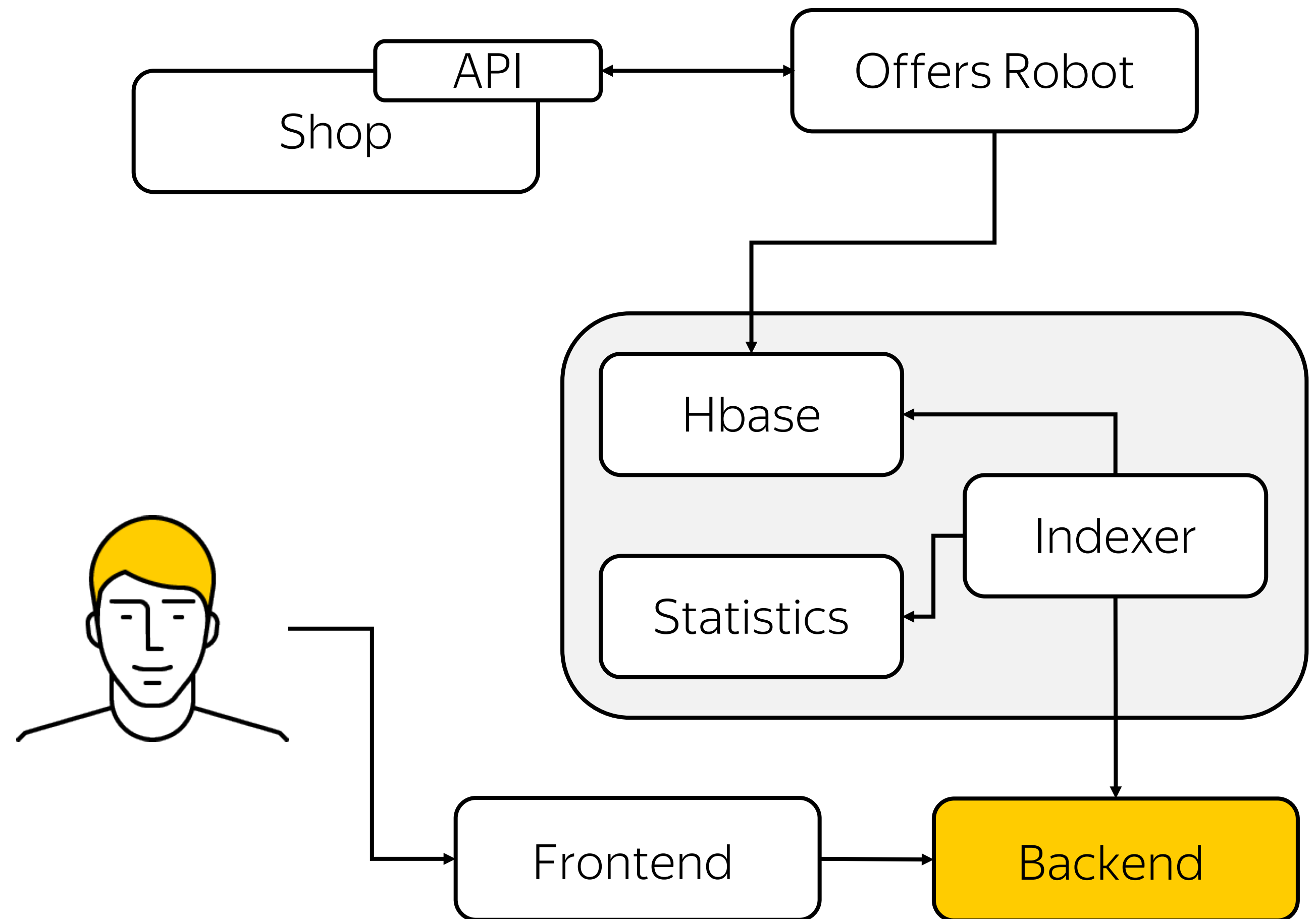
Яндекс.Маркет – построение индекса

Магазины экспортируют список предложений в API

Offers robot собирает предложения в hbase

Indexer вычитывает таблицу и строит индекс

Выкладка индекса



Система выкладки



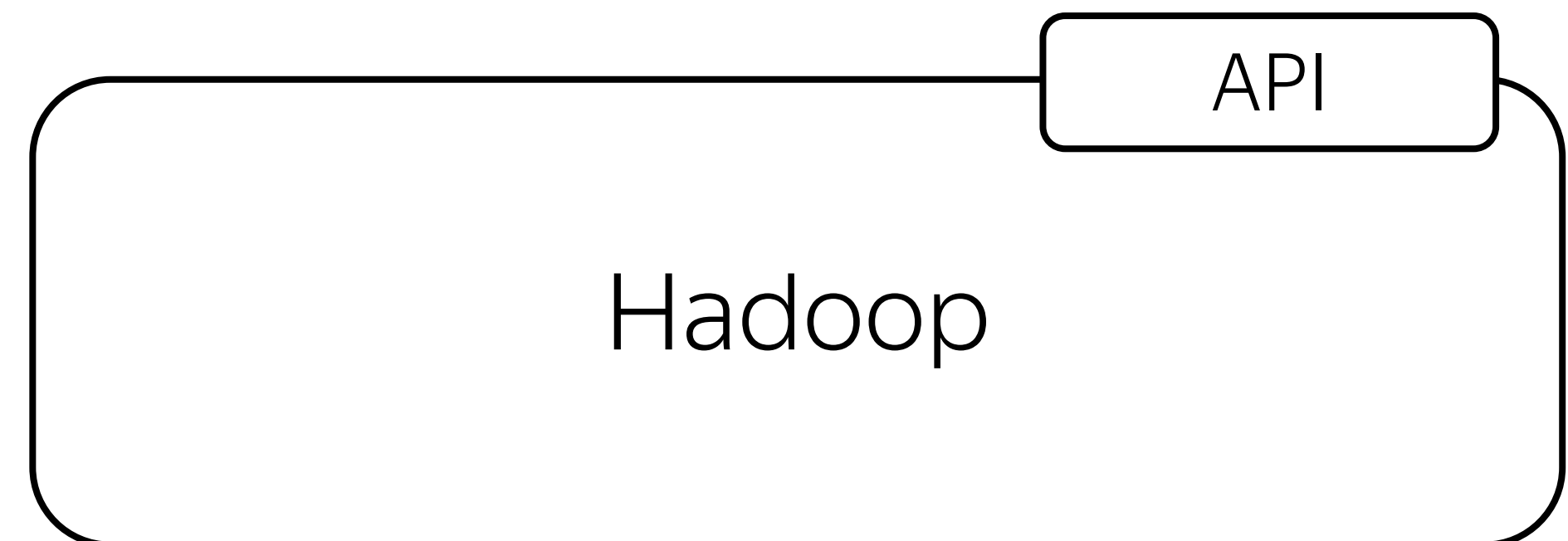
Hadoop as a Service

Hadoop SRE:

- › Разрабатывают и управляют платформой

Service SRE

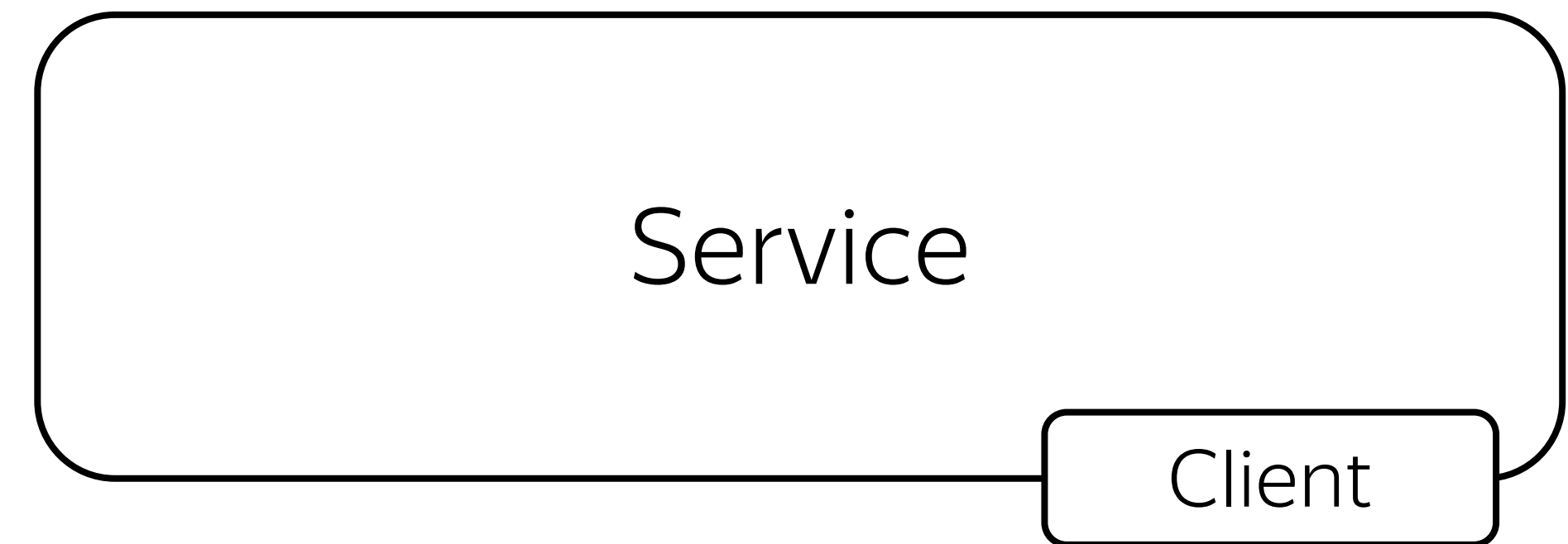
Hadoop SRE



Hadoop as a Service

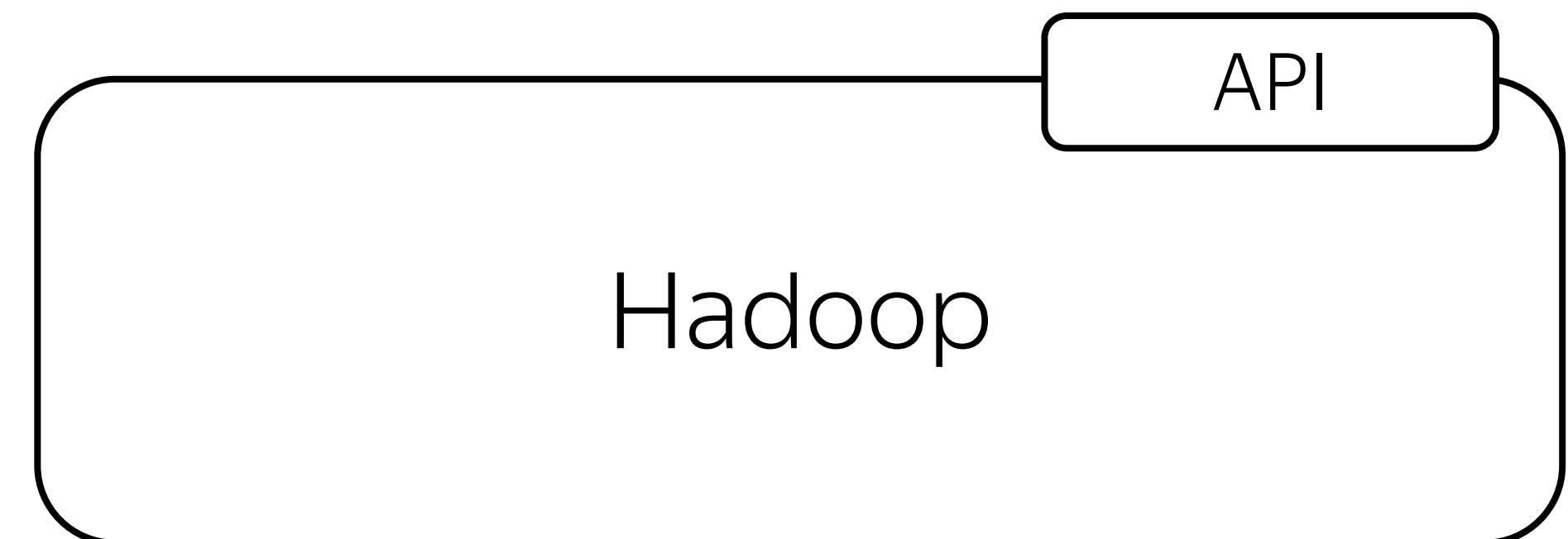
Hadoop SRE:

- › Разрабатывают и управляют платформой
- › Приложения - зона ответственности заказчика



Service SRE

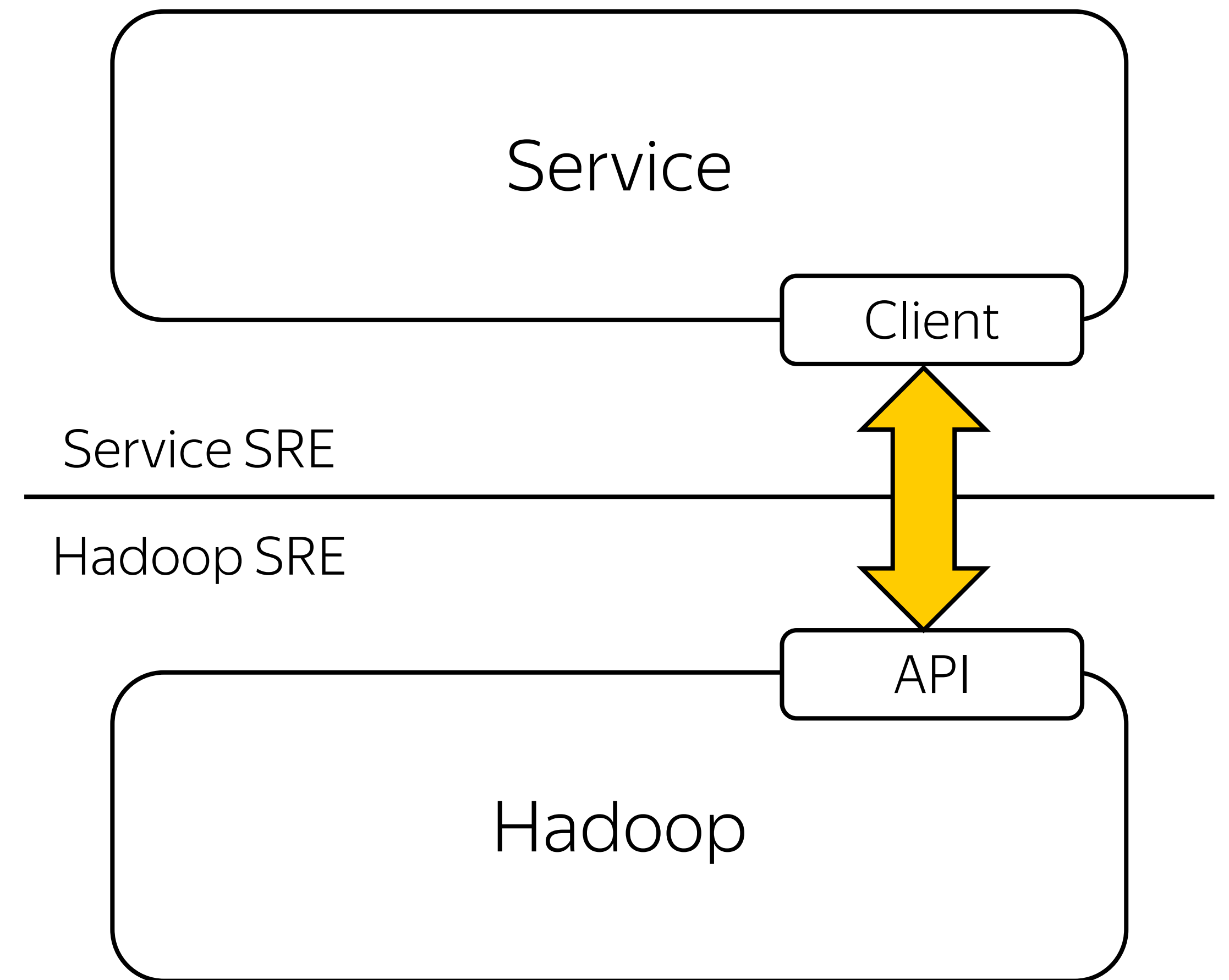
Hadoop SRE



Hadoop as a Service

Hadoop SRE:

- › Разрабатывают и управляют платформой
- › Приложения - зона ответственности заказчика



Зачем нужна своя сборка Hadoop

- Отсутствует vendor-lock

Зачем нужна своя сборка Hadoop

- Отсутствует vendor-lock
- Возможность развивать платформу под запросы сервисов

Зачем нужна своя сборка Hadoop

- Отсутствует vendor-lock
- Возможность развивать платформу под запросы сервисов
- Команда разработки и поддержки

Зачем нужна своя сборка Hadoop

- Отсутствует vendor-lock
- Возможность развивать платформу под запросы сервисов
- Команда разработки и поддержки
- Необходимость адаптации под собственную инфраструктуру

Наша сборка Hadoop

Основана на бинарных пакетах

Наша сборка Hadoop

- Основана на бинарных пакетах
- Собственные патчи и изменения

Наша сборка Hadoop

- Основана на бинарных пакетах
- Собственные патчи и изменения
- Работает на выделенных серверах

Наша сборка Hadoop

- Основана на бинарных пакетах
- Собственные патчи и изменения
- Работает на выделенных серверах
- Разворачивается с помощью saltstack

Наша сборка Hadoop

- Основана на бинарных пакетах
- Собственные патчи и изменения
- Работает на выделенных серверах
- Разворачивается с помощью saltstack
- Конфигурация хранится в git

Salt

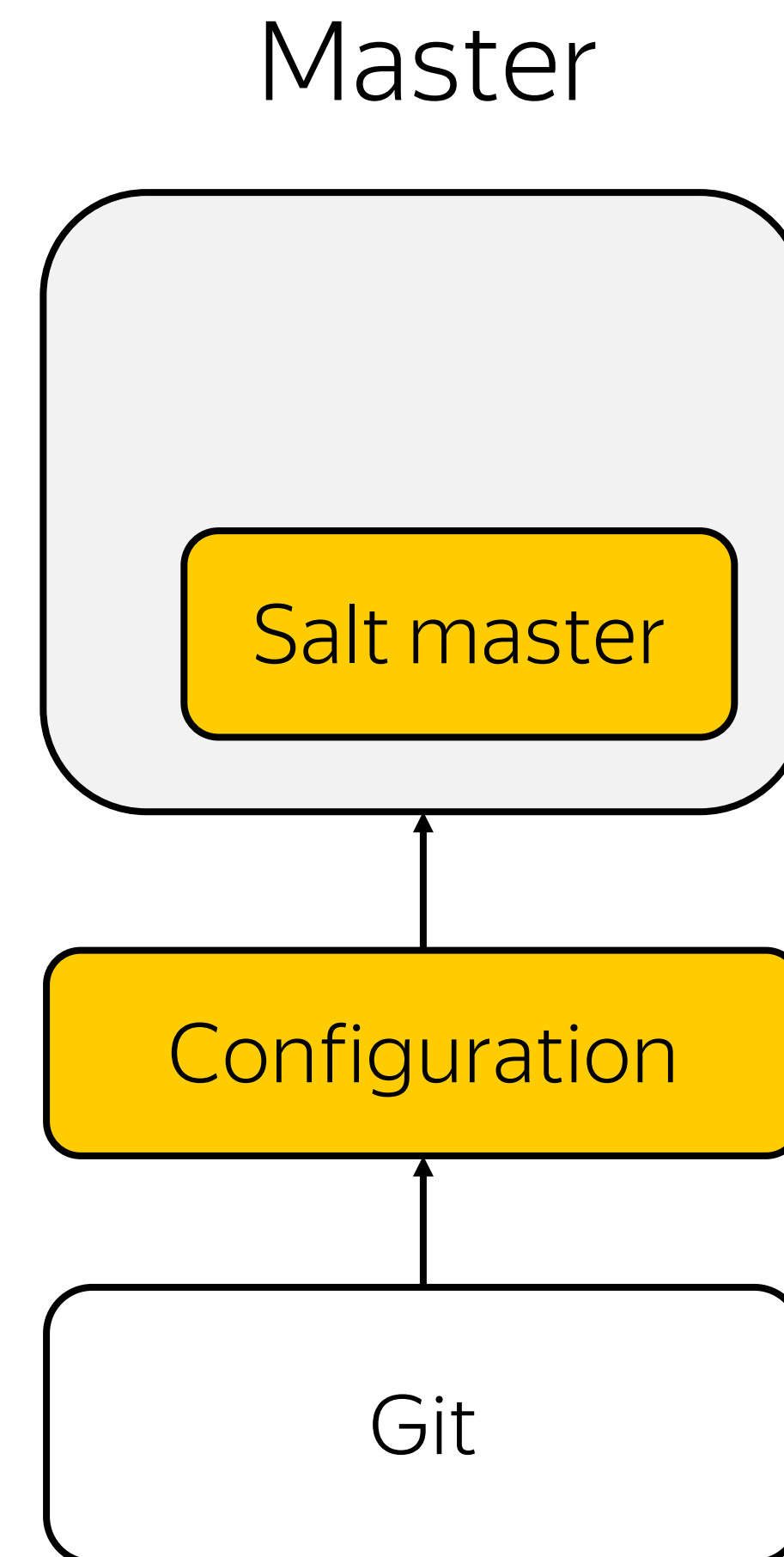
Выделенный salt-master

Master



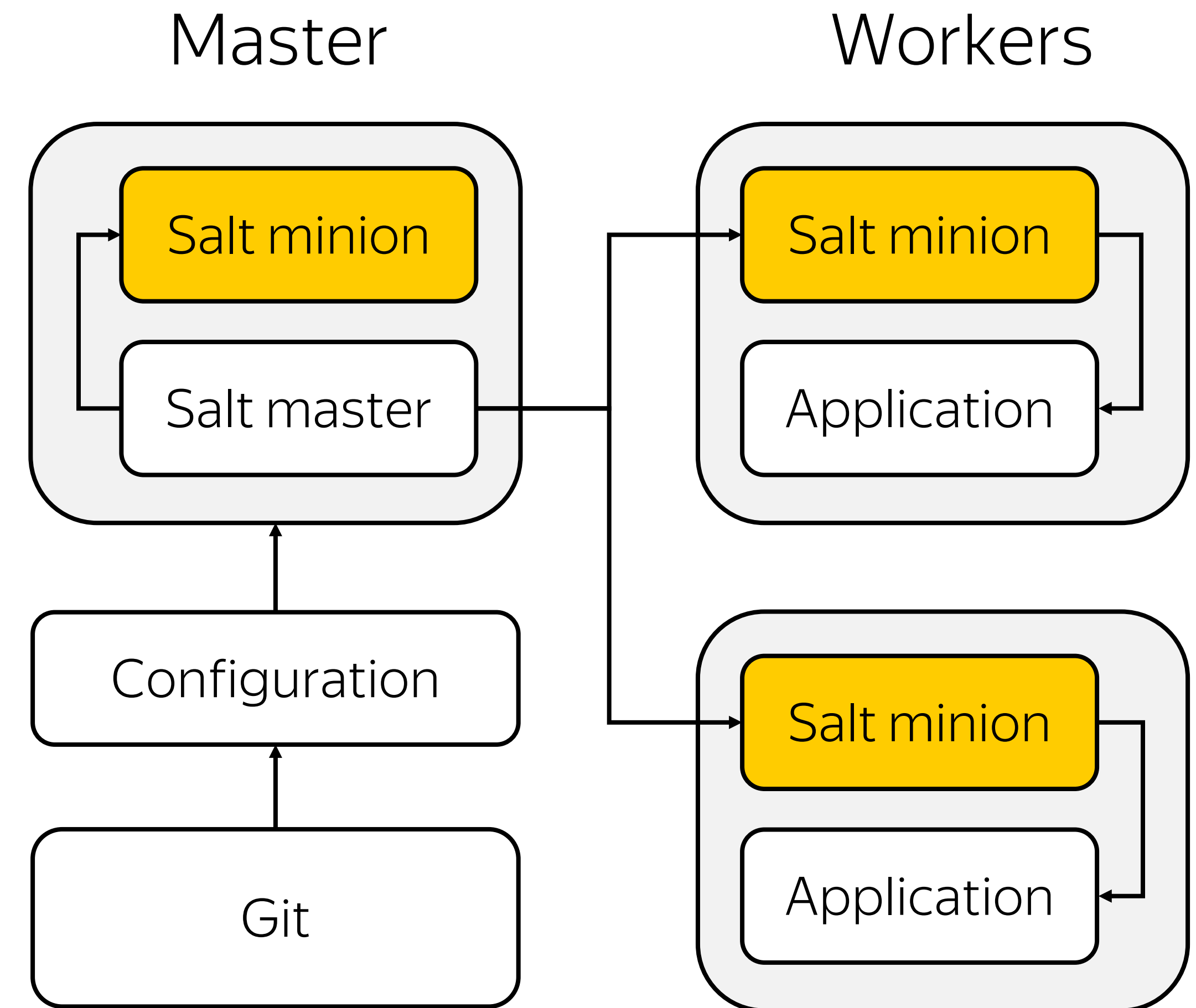
Salt

- Выделенный salt-master
- Конфигурация в git



Salt

- Выделенный salt-master
- Конфигурация в git
- Salt-minion на каждой машине



Мониторинги

Своя система мониторинга juggler

› <https://events.yandex.ru/lib/talks/35>

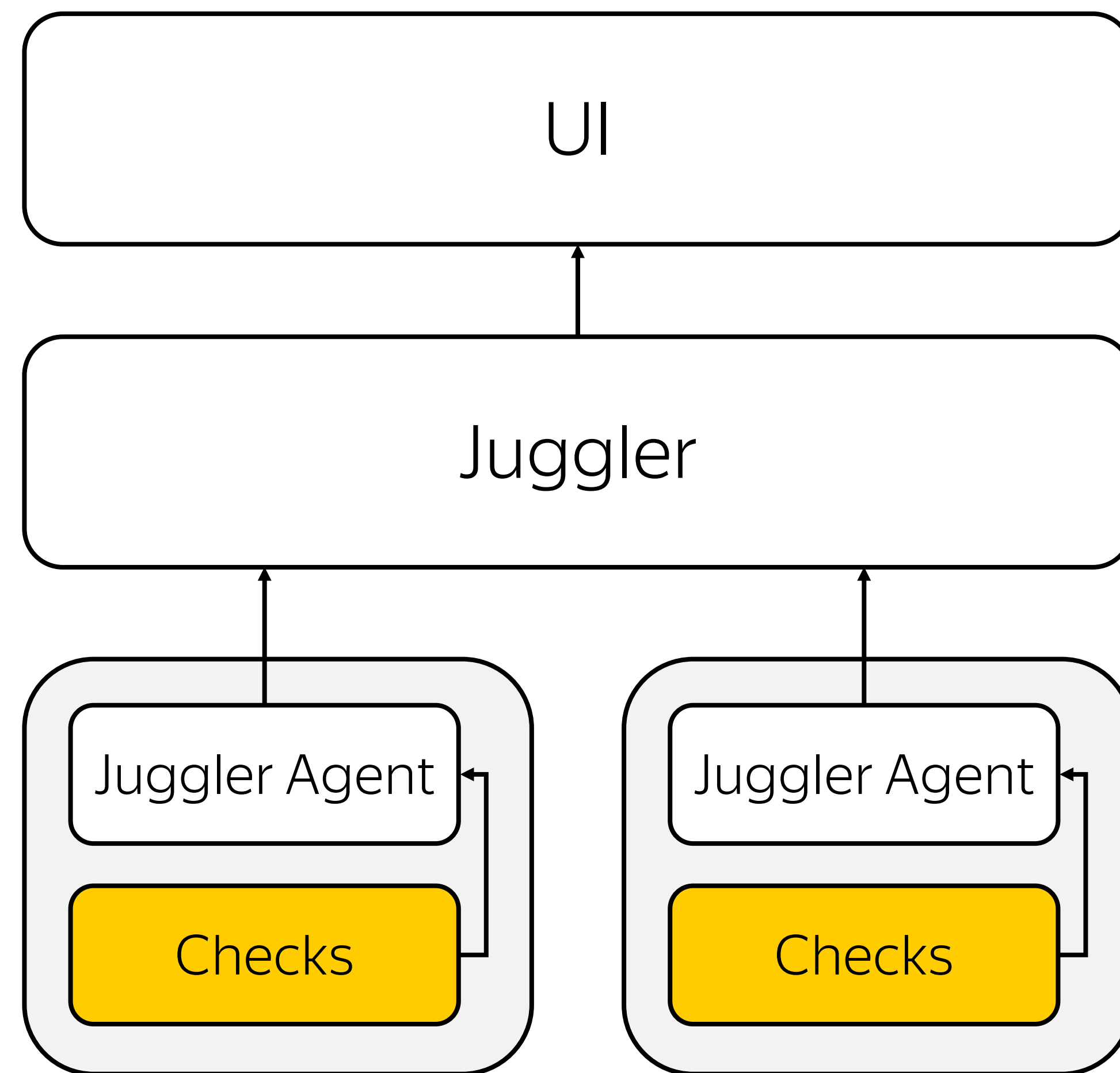
Мониторинги

Своя система мониторинга juggler

› <https://events.yandex.ru/lib/talks/35>

Набор проверок:

- › Base: cpu, mem, space, network, hw_errs, ...
- › Services
- › Integration



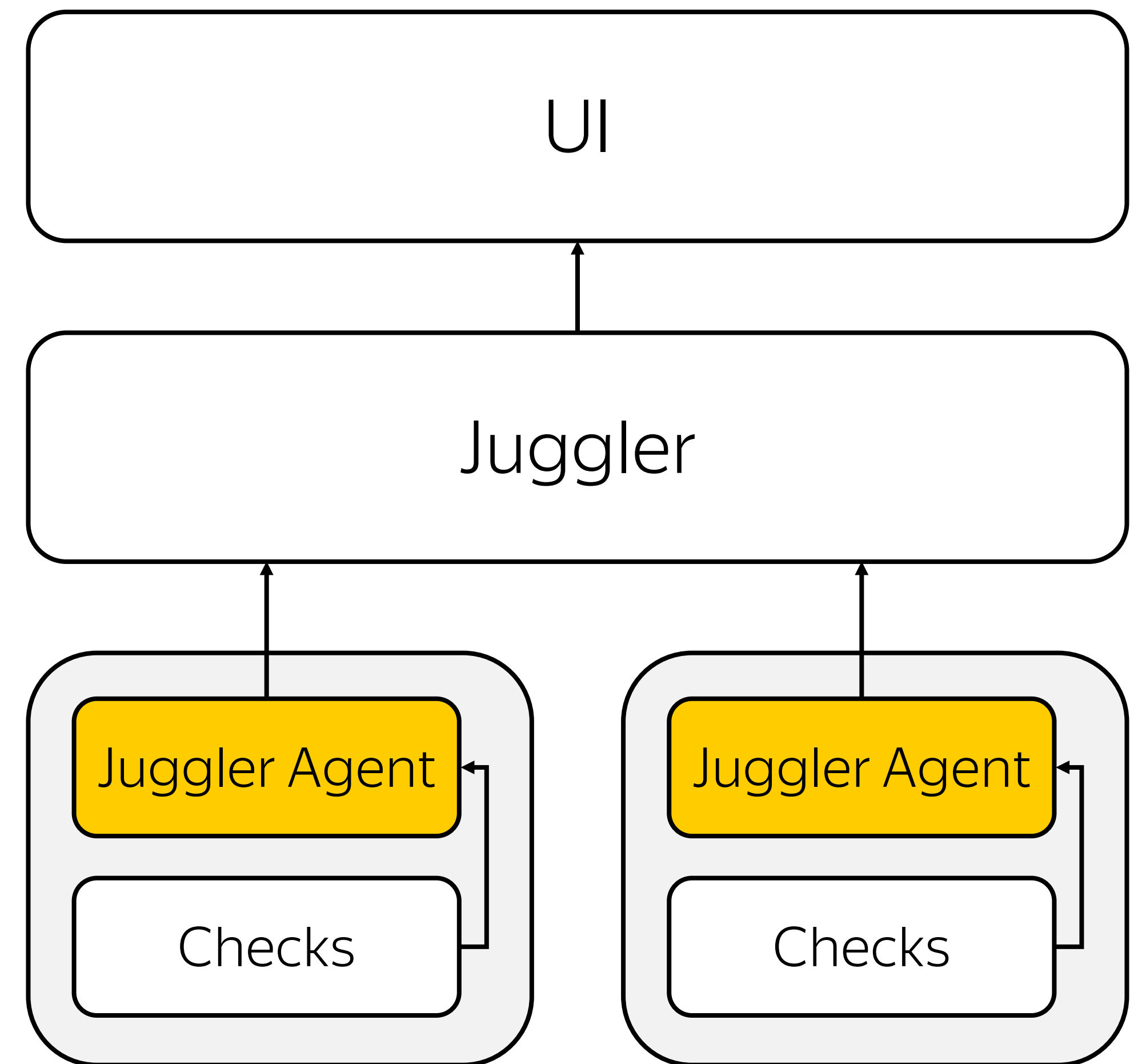
Мониторинги

Своя система мониторинга juggler

› <https://events.yandex.ru/lib/talks/35>

Набор проверок:

- › Base: cpu, mem, space, network, hw_errs, ...
- › Services
- › Integration



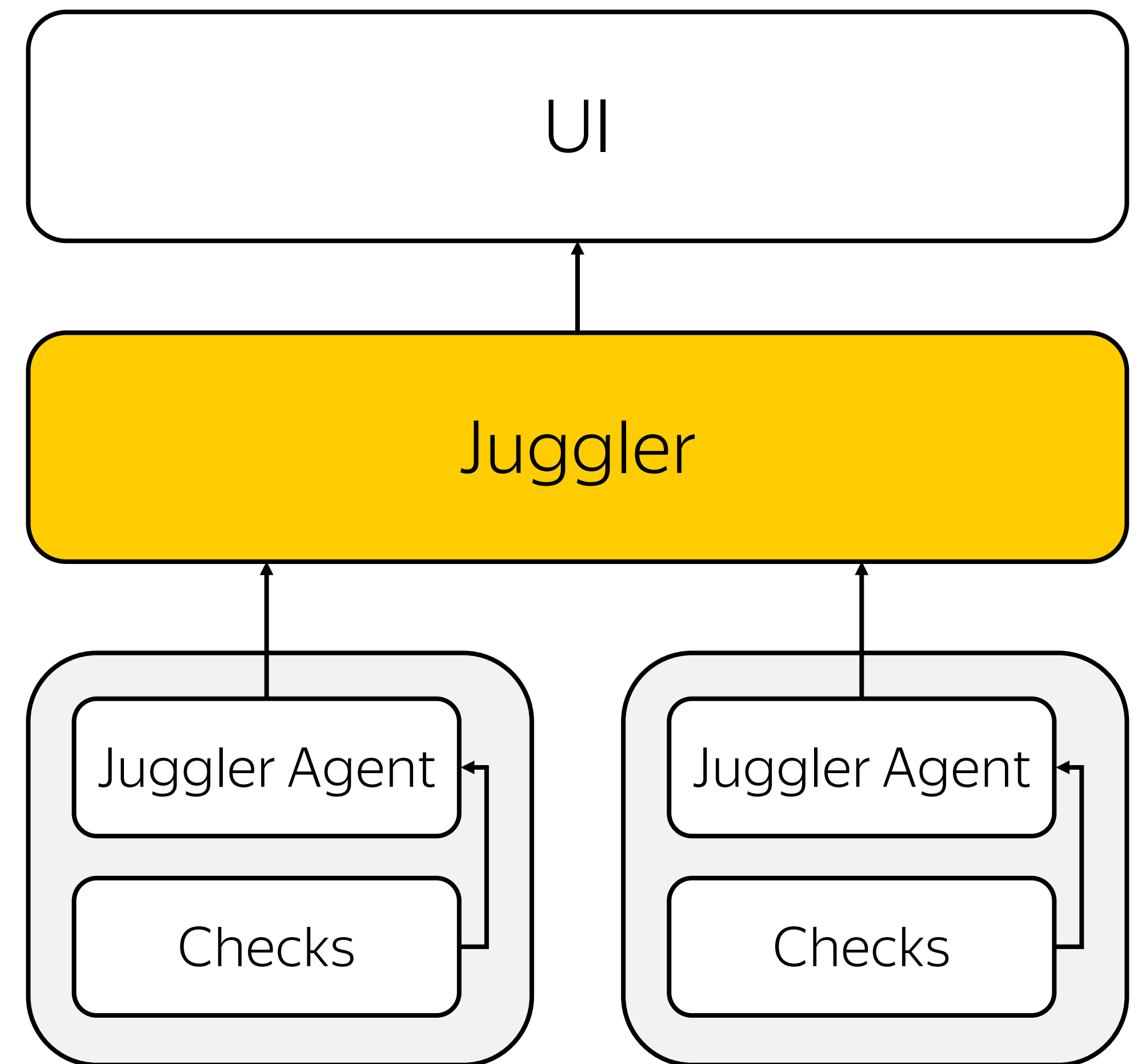
Мониторинги

Своя система мониторинга juggler

› <https://events.yandex.ru/lib/talks/35>

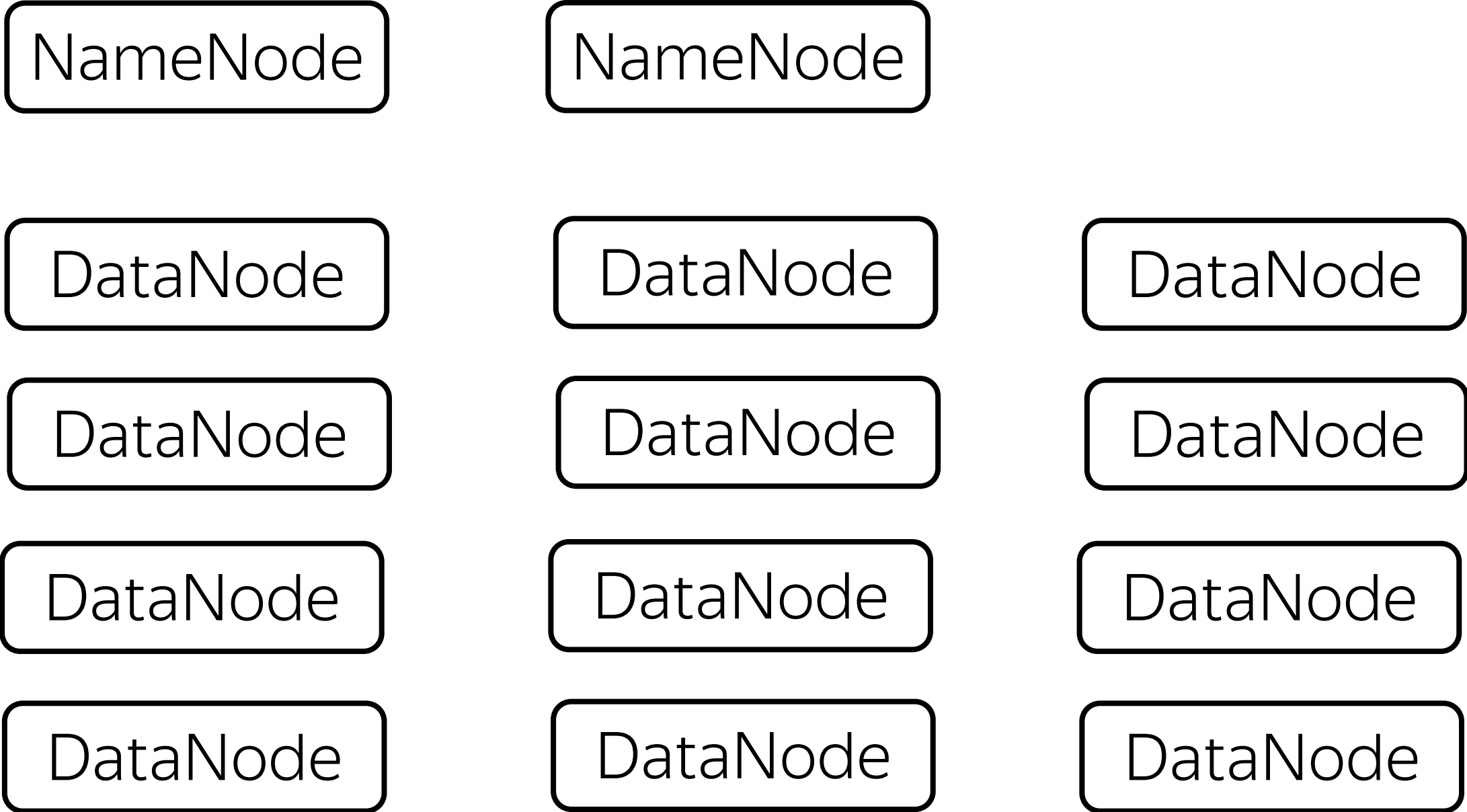
Набор проверок:

- › Base: cpu, mem, space, network, hw_errs, ...
- › Services
- › Integration



Мониторинги - агрегация

Виртуальные хосты для агрегации

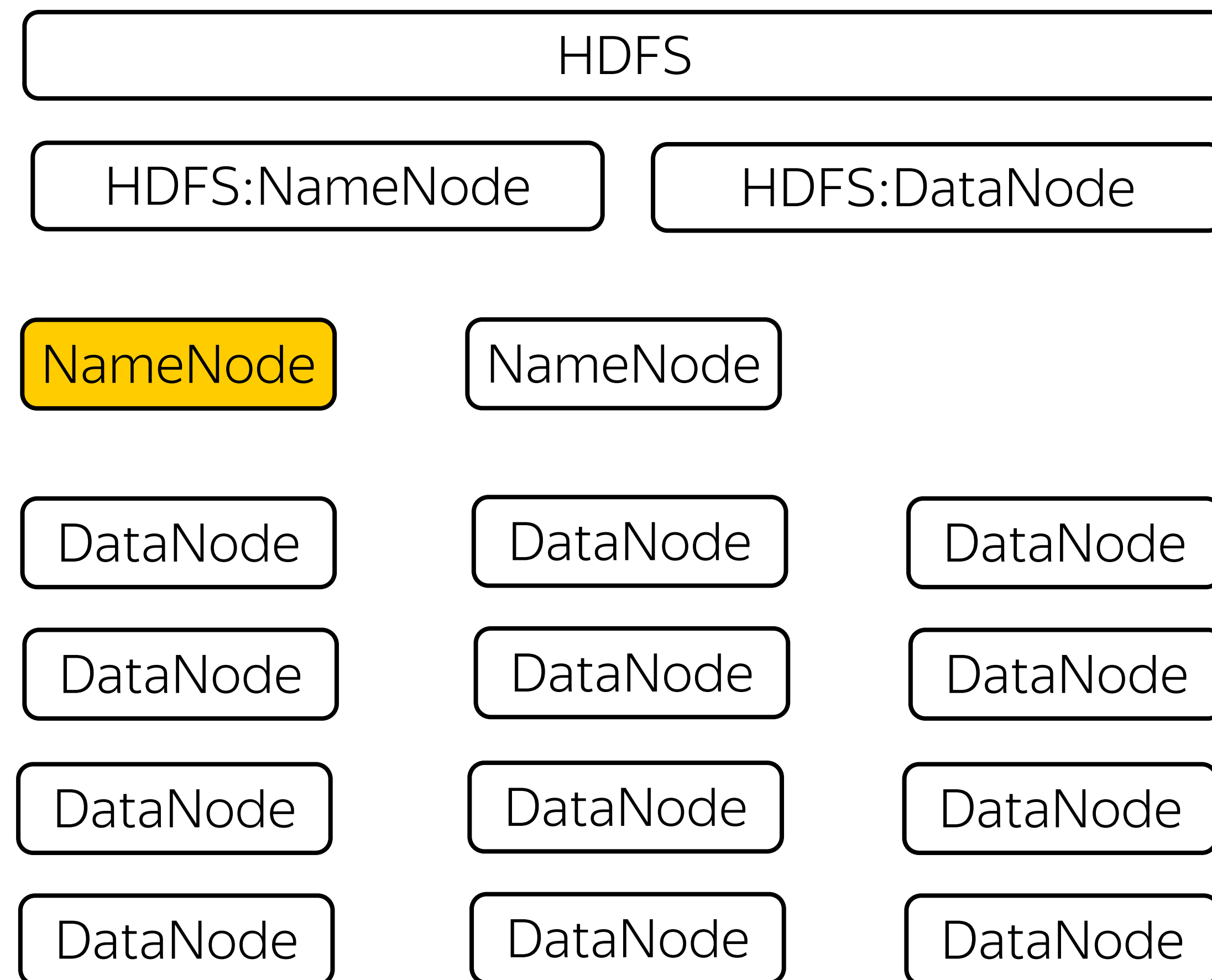


Мониторинги - агрегация

Виртуальные хосты для агрегации

Кластер:

› По нему звонят по ночам

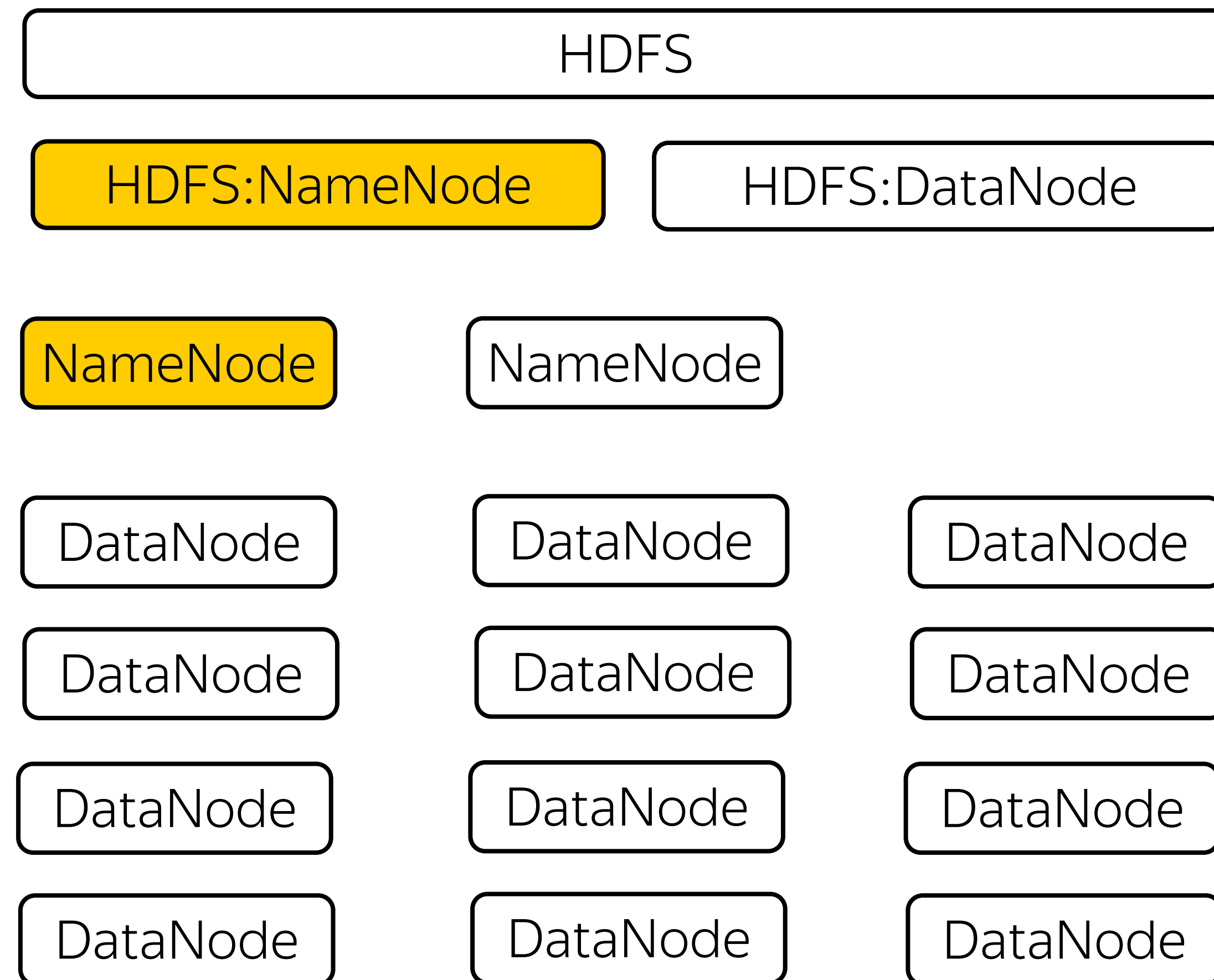


Мониторинги - агрегация

Виртуальные хосты для агрегации

Кластер:

- › По нему звонят по ночам
- › Зажигаются только критичные для работоспособности кластера проверки

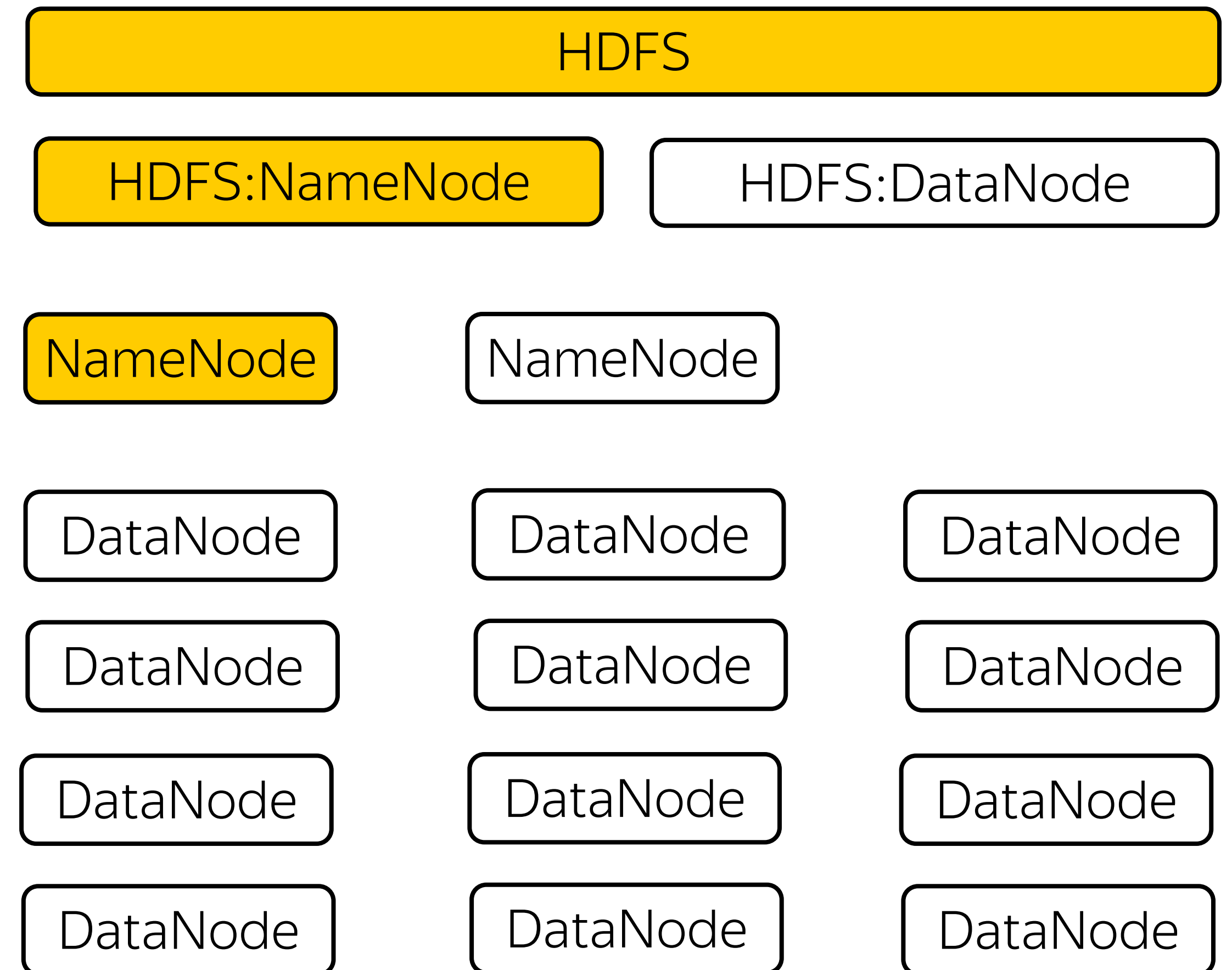


Мониторинги - агрегация

Виртуальные хосты для агрегации

Кластер:

- › По нему звонят по ночам
- › Зажигаются только критичные для работоспособности кластера проверки

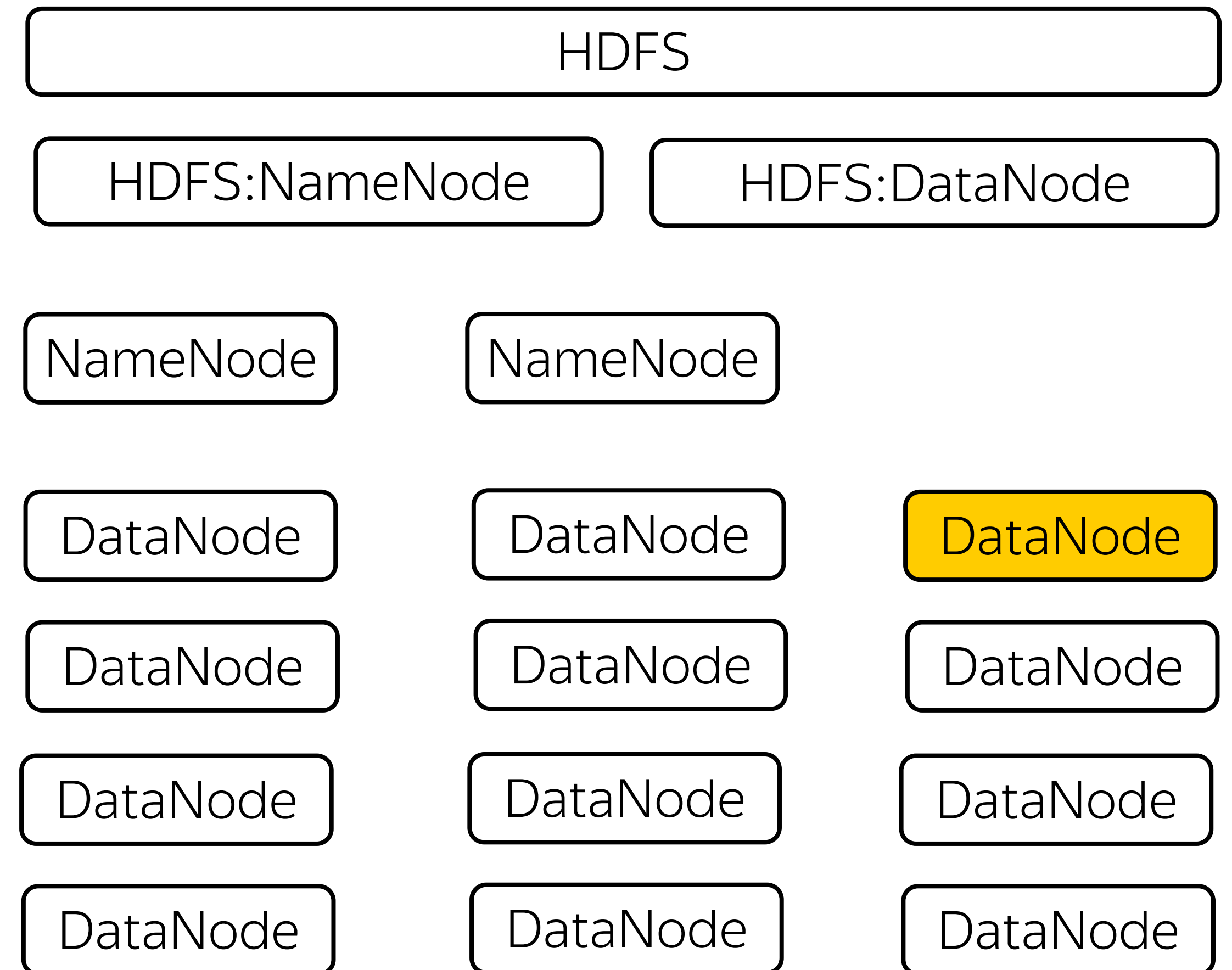


Мониторинги - агрегация

Виртуальные хосты для агрегации

Кластер:

- › По нему звонят по ночам
- › Зажигаются только критичные для работоспособности кластера проверки
- › Порог срабатывания для обычных проверок

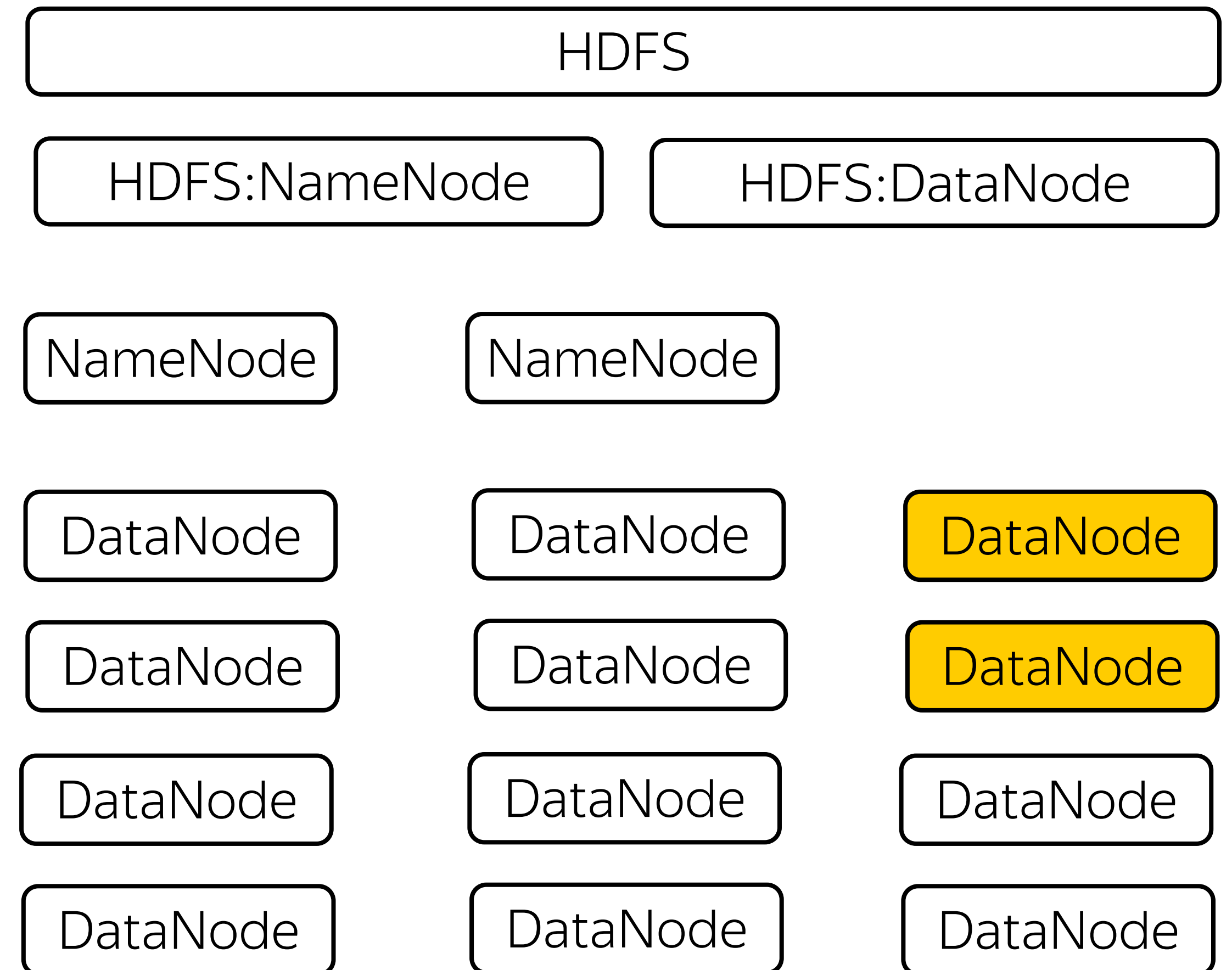


Мониторинги - агрегация

Виртуальные хосты для агрегации

Кластер:

- › По нему звонят по ночам
- › Зажигаются только критичные для работоспособности кластера проверки
- › Порог срабатывания для обычных проверок

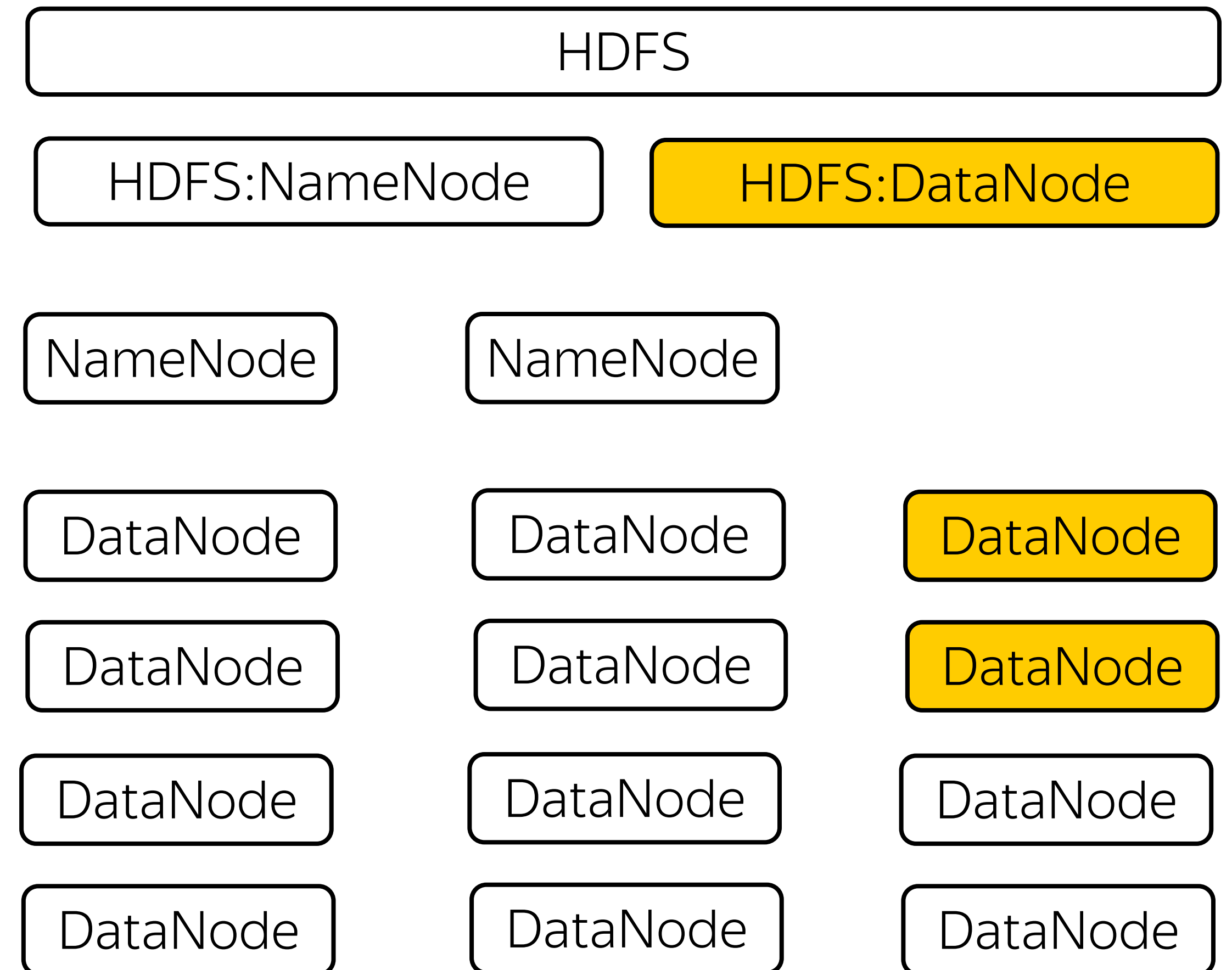


Мониторинги - агрегация

Виртуальные хосты для агрегации

Кластер:

- › По нему звонят по ночам
- › Зажигаются только критичные для работоспособности кластера проверки
- › Порог срабатывания для обычных проверок

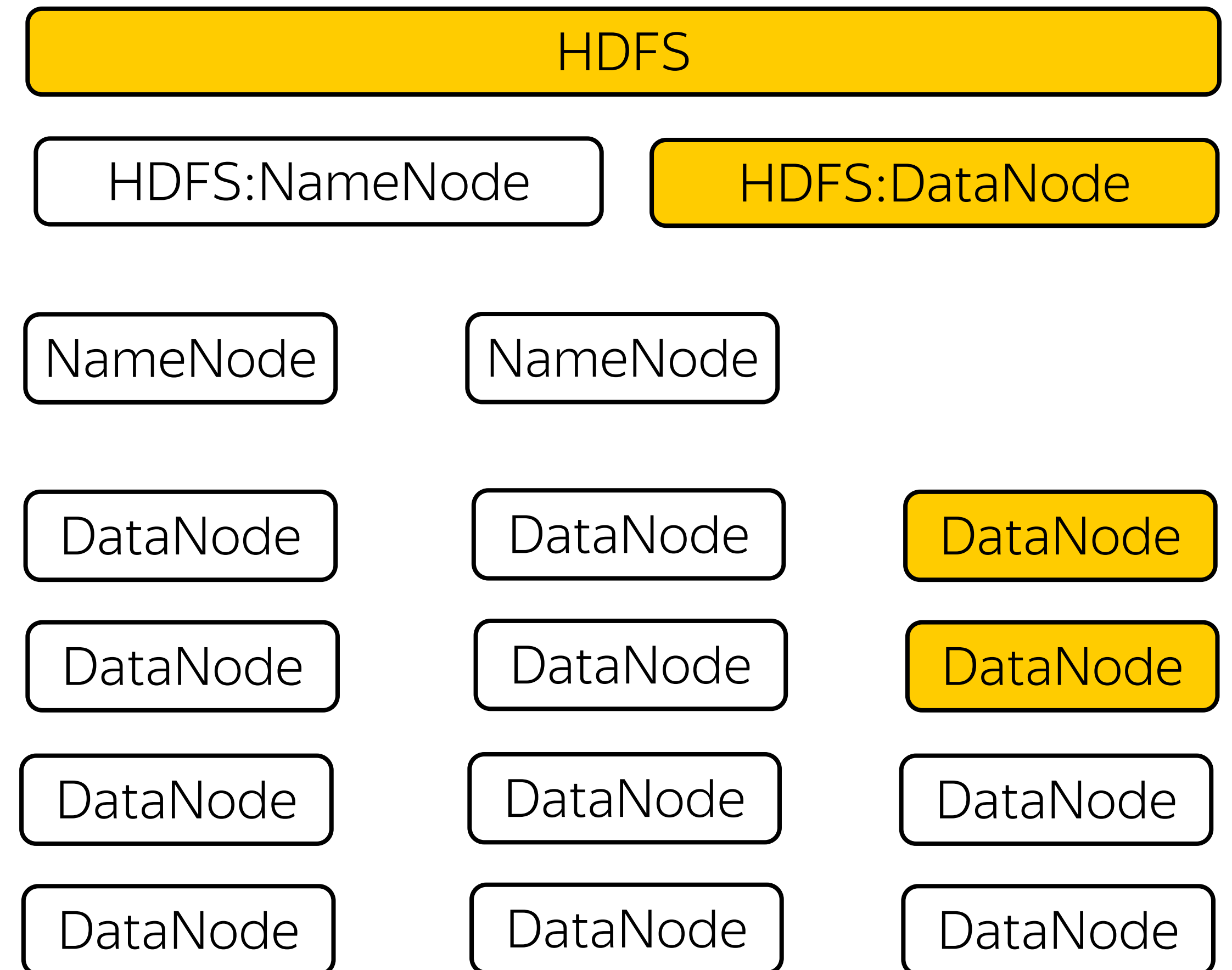


Мониторинги - агрегация

Виртуальные хосты для агрегации

Кластер:

- › По нему звонят по ночам
- › Зажигаются только критичные для работоспособности кластера проверки
- › Порог срабатывания для обычных проверок

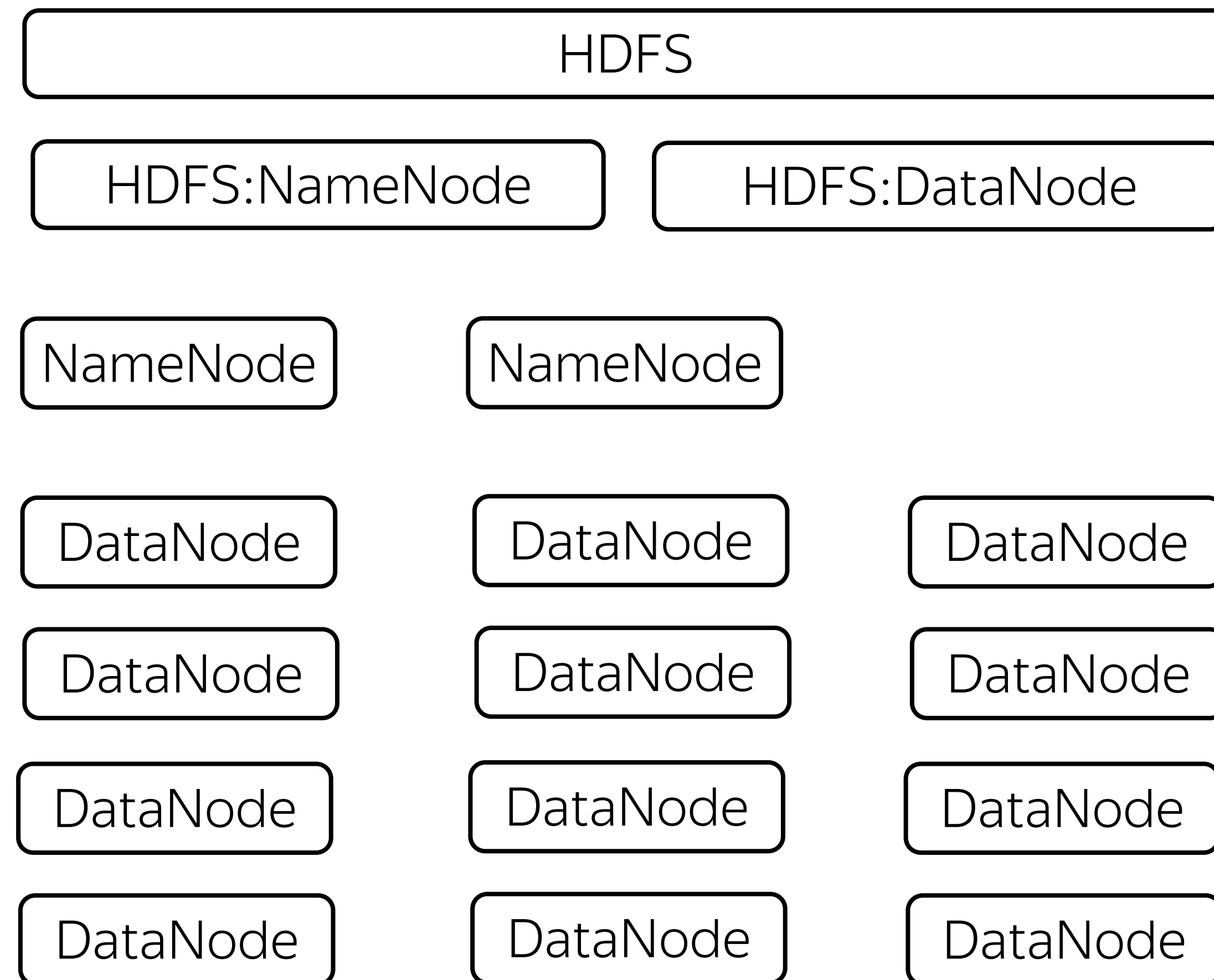


Мониторинги - агрегация

Виртуальные хосты для агрегации

Машины кластера:

- › По ним не звонят
- › Зажигаются все проверки
- › Нет порогов срабатывания

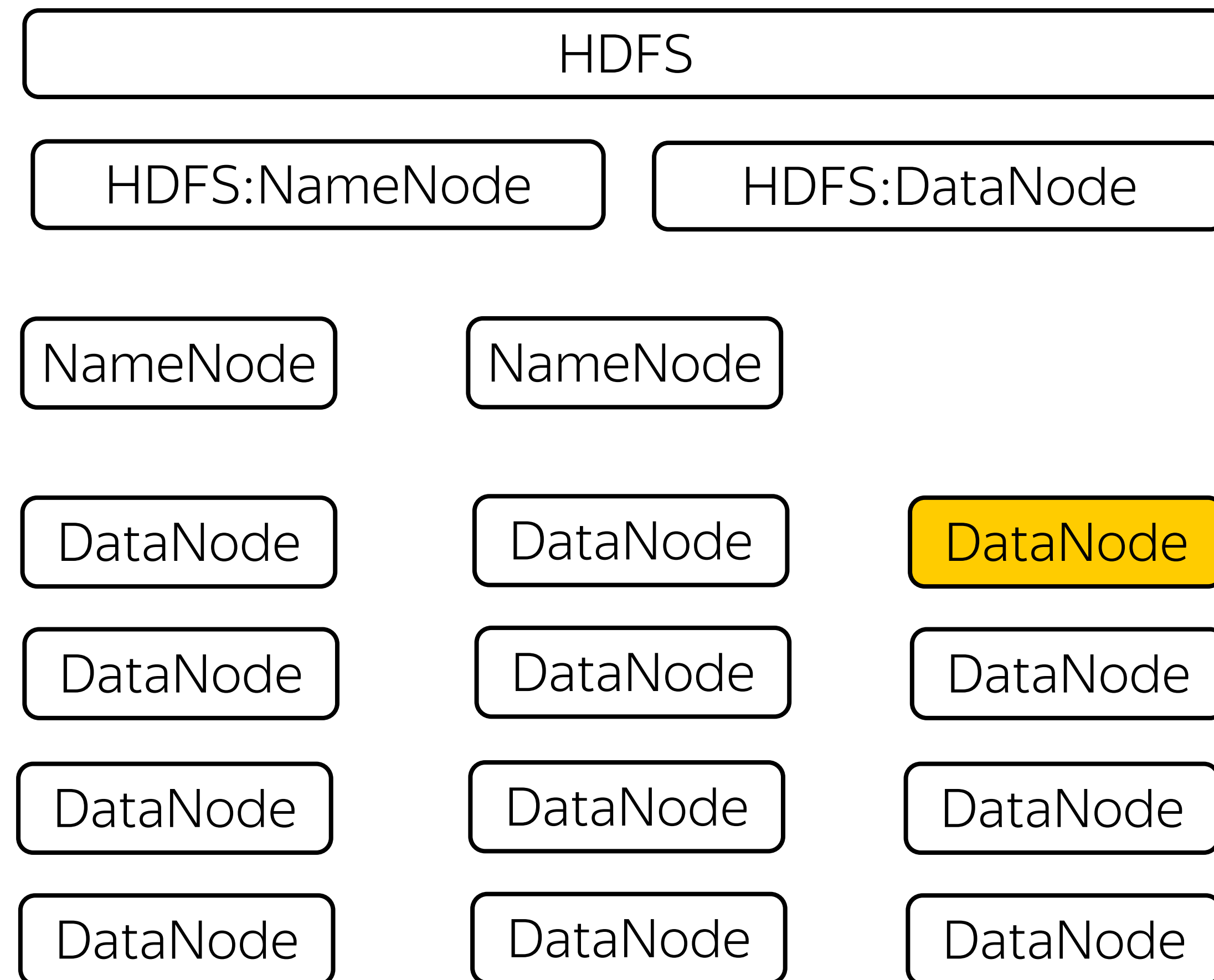


Мониторинги - агрегация

Виртуальные хосты для агрегации

Машины кластера:

- › По ним не звонят
- › Зажигаются все проверки
- › Нет порогов срабатывания

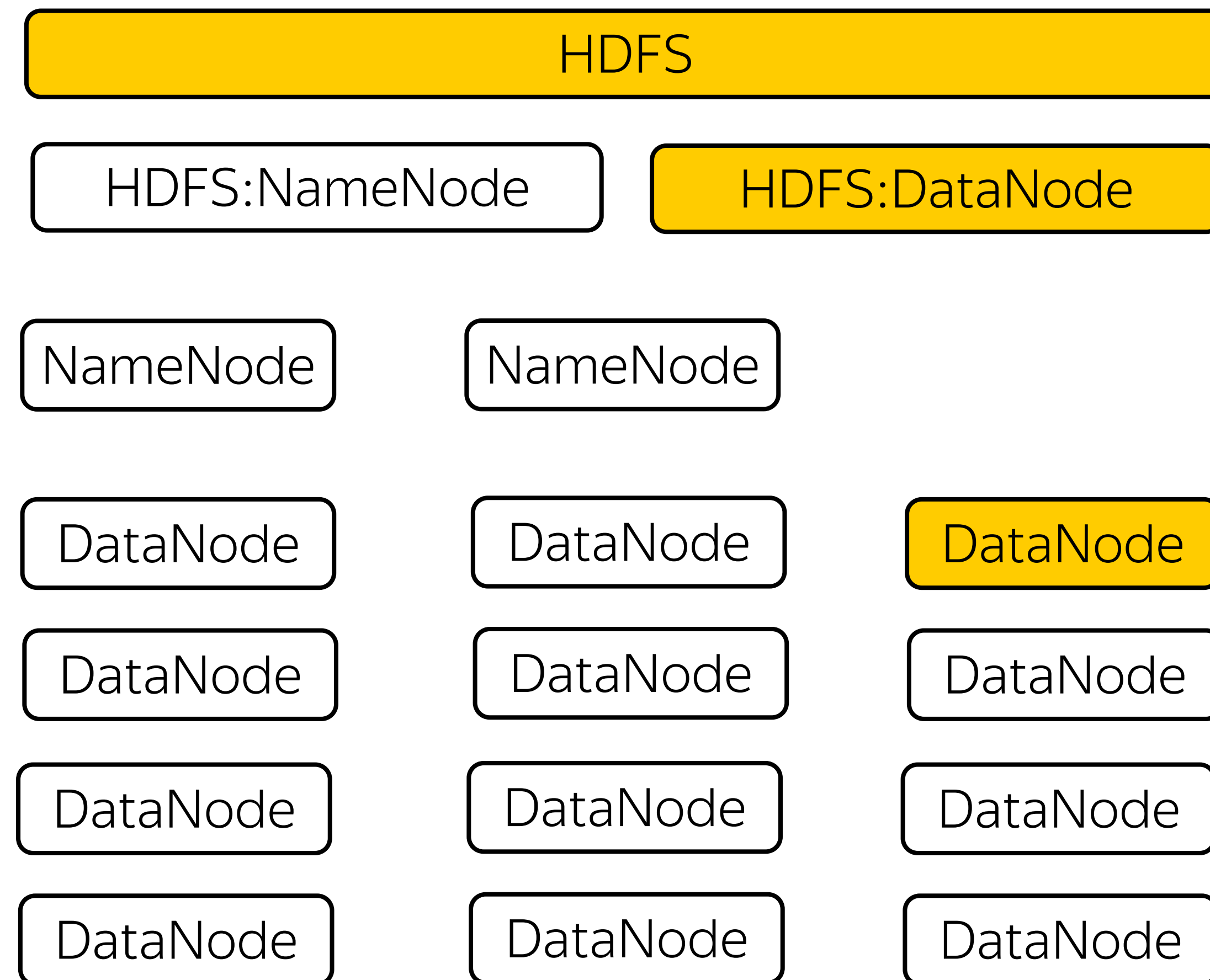


Мониторинги - агрегация

Виртуальные хосты для агрегации

Машины кластера:

- › По ним не звонят
- › Зажигаются все проверки
- › Нет порогов срабатывания



Мониторинги - CI

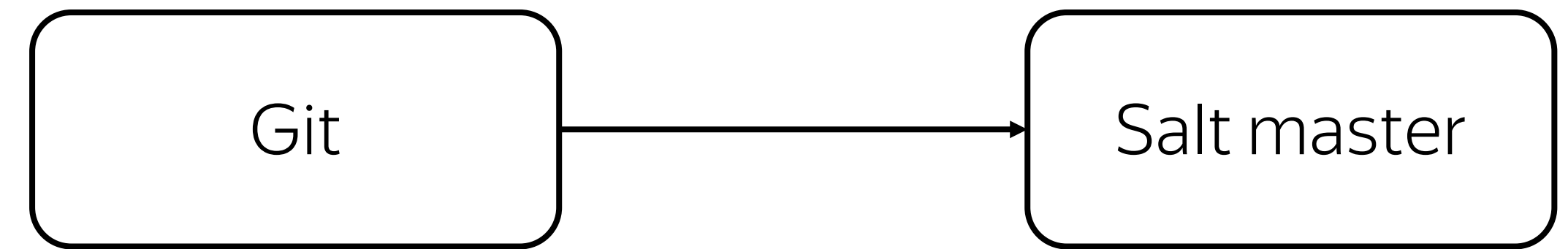
Агрегация по тегам

- › Проверки по группам машин с определенным тегом

Мониторинги - CI

Агрегация по тегам

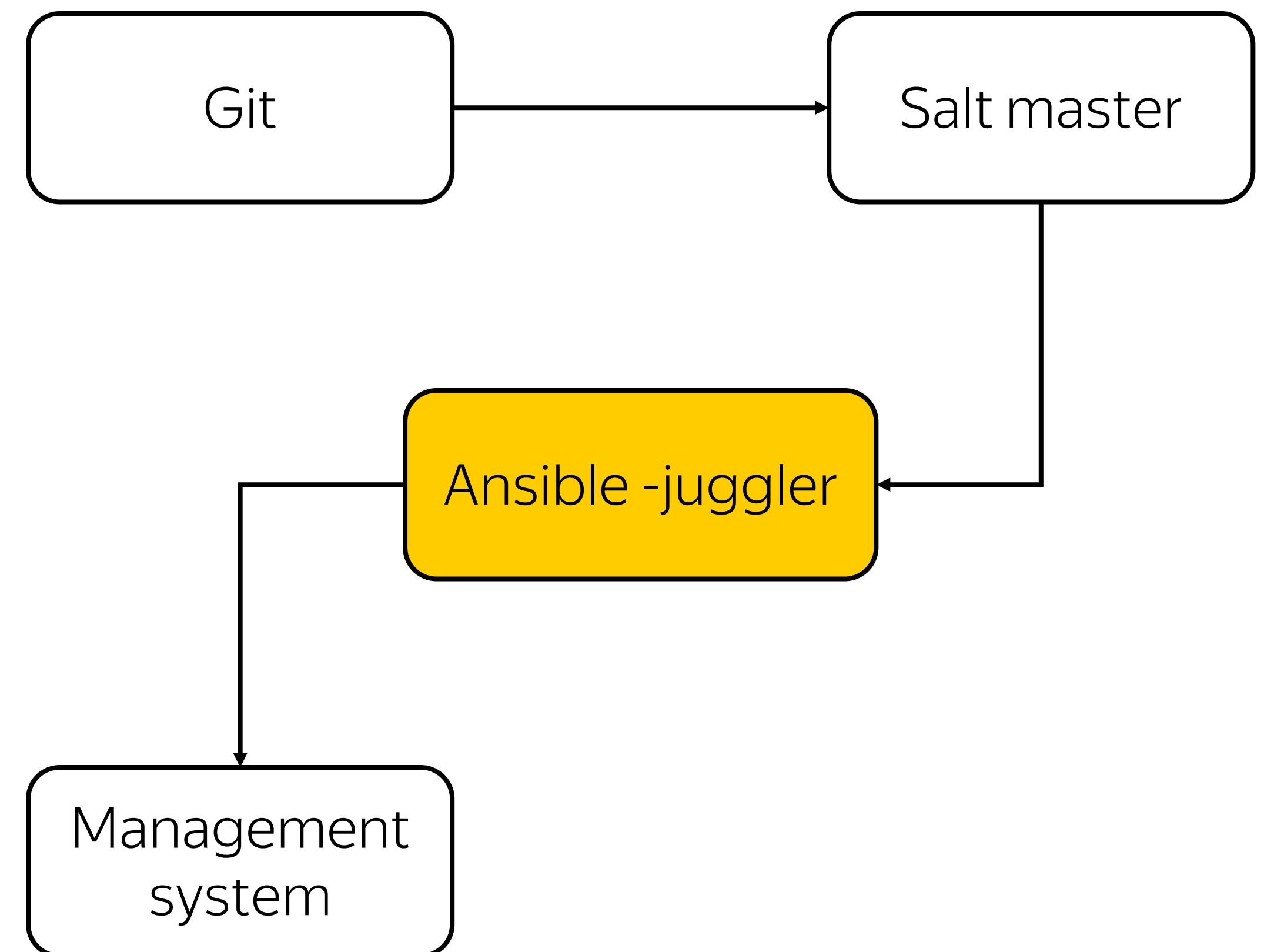
- › Проверки по группам машин с определенным тегом
- › Теги строятся из конфигурации кластера



Мониторинги - CI

Агрегация по тегам

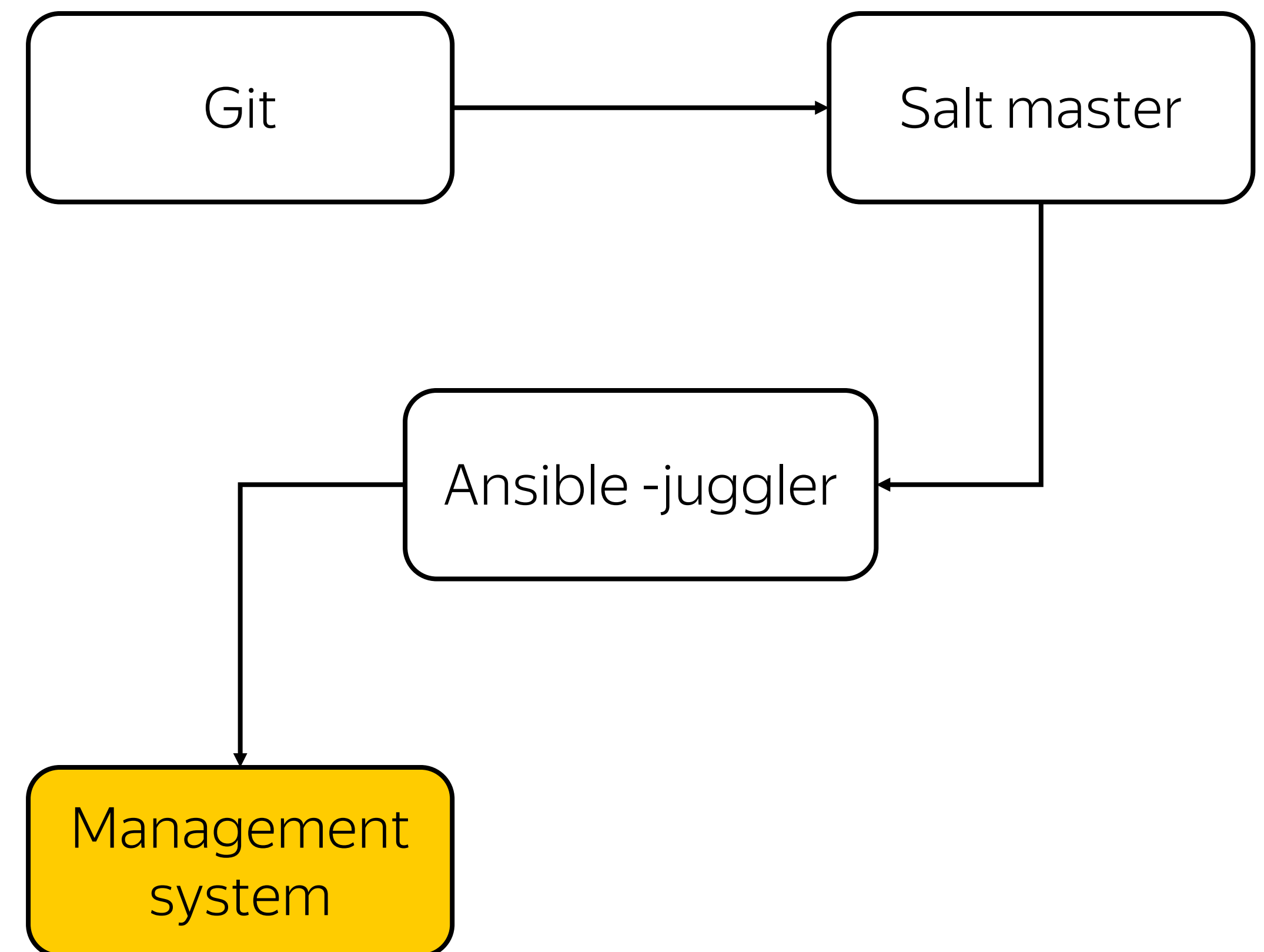
- › Проверки по группам машин с определенным тегом
- › Теги строятся из конфигурации кластера
- › При изменении кластера - автоматически обновляем теги



Мониторинги - CI

Агрегация по тегам

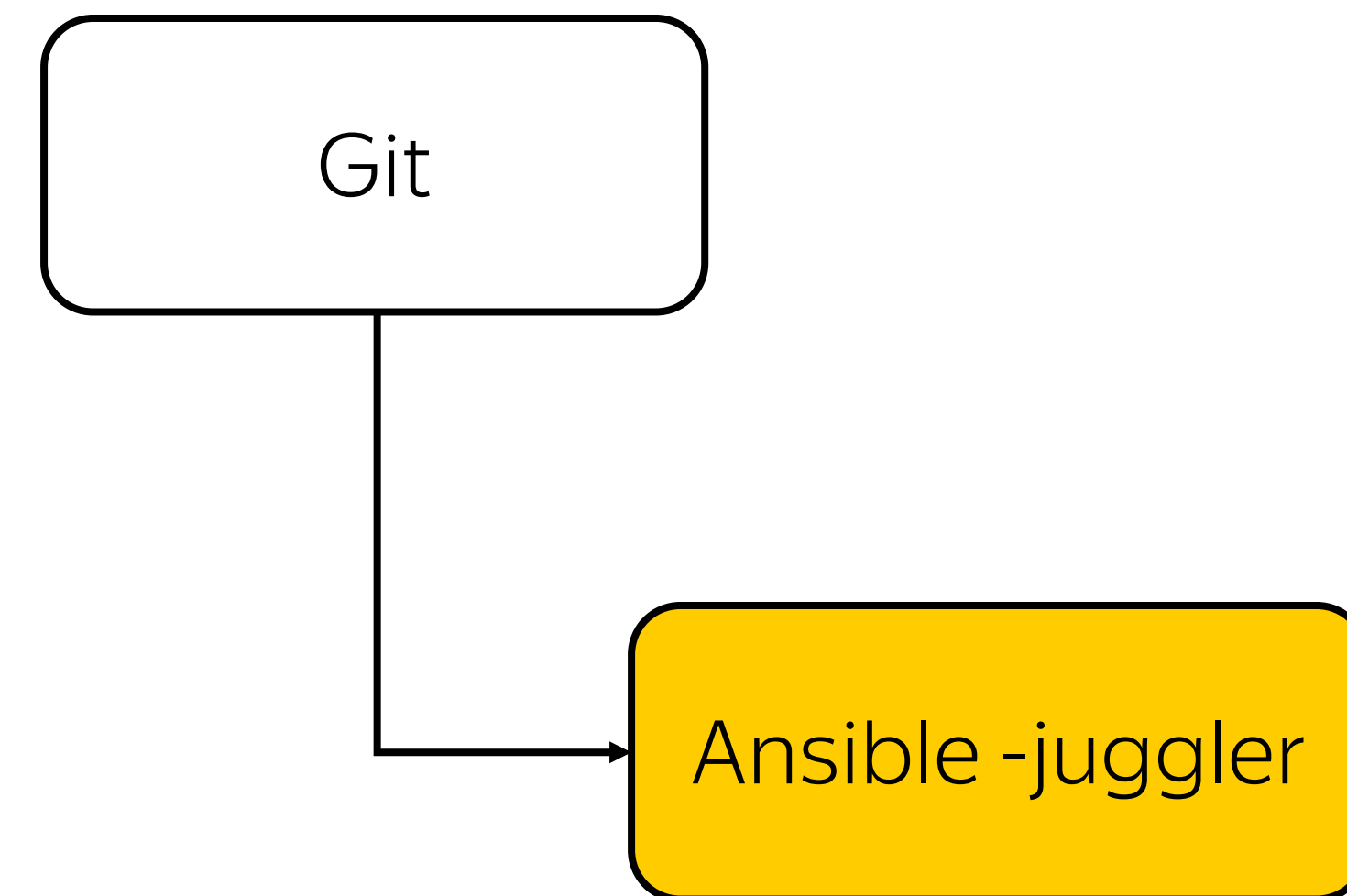
- › Проверки по группам машин с определенным тегом
- › Теги строятся из конфигурации кластера
- › При изменении кластера - автоматически обновляем теги



Мониторинги - CD

Ansible-juggler

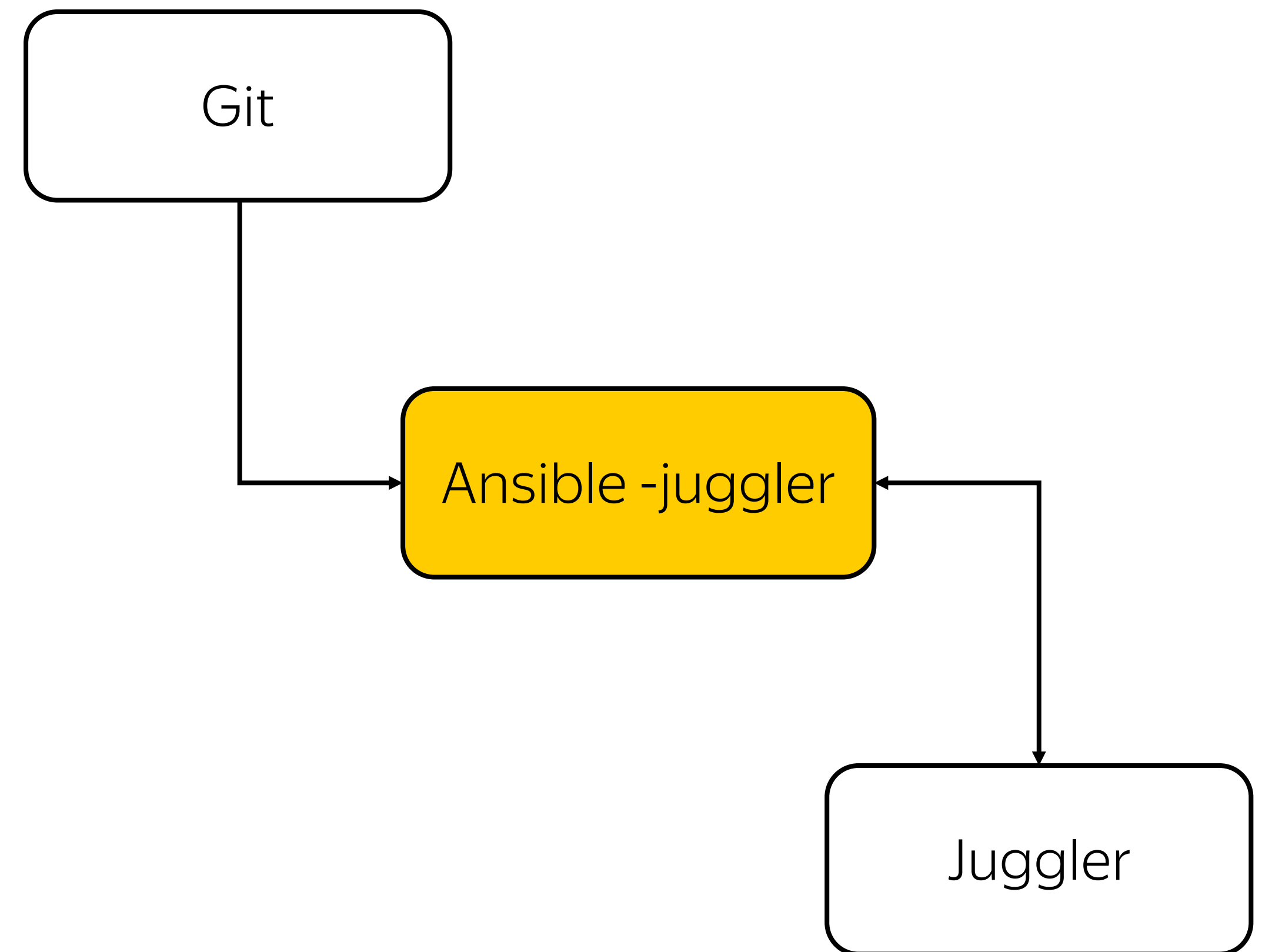
- › Читаемый yaml-file
- › Правила агрегации и время срабатывания проверок



Мониторинги - CD

Ansible-juggler

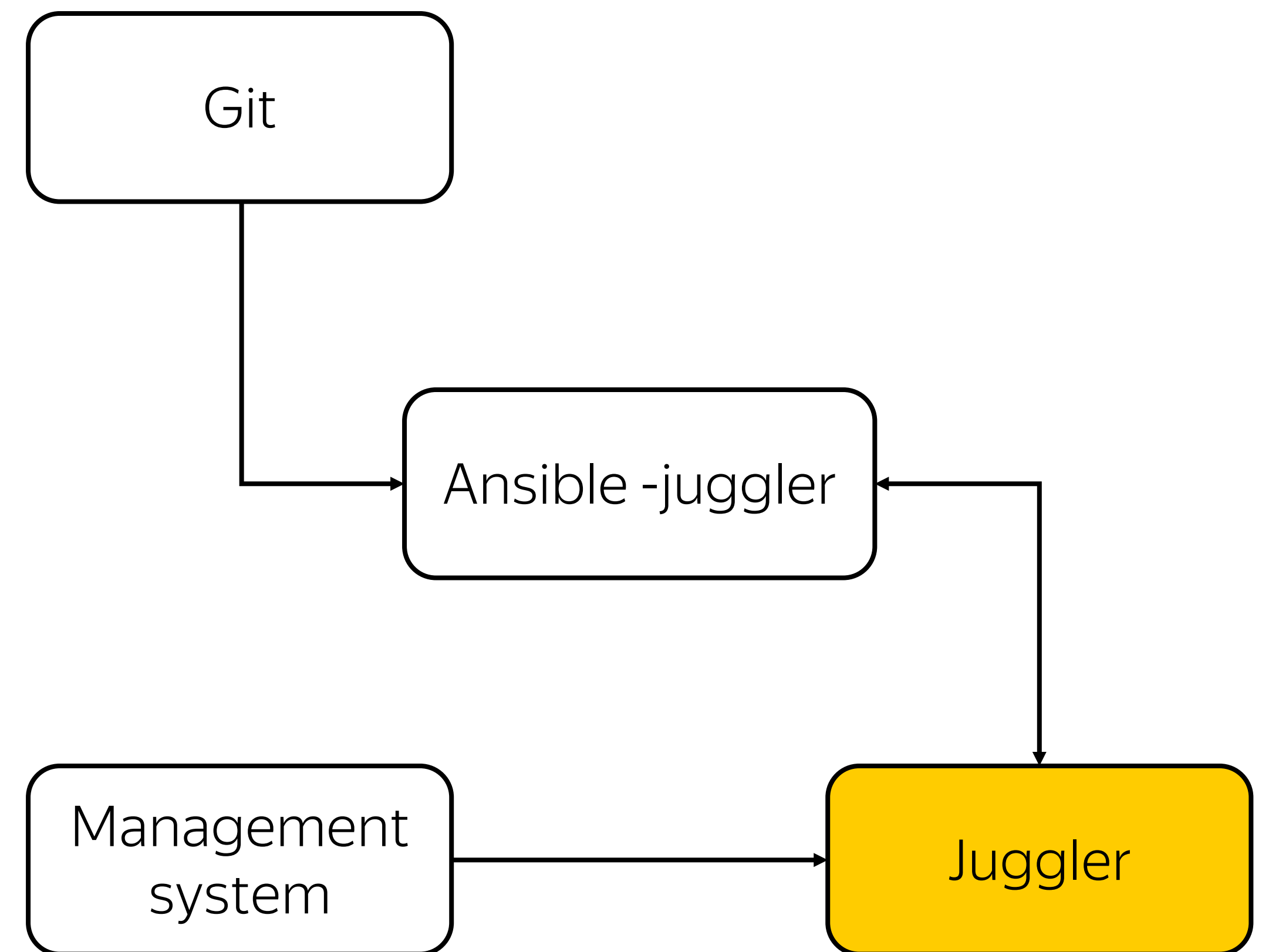
- › Читаемый yaml-file
- › Правила агрегации и время срабатывания проверок
- › Актуализирует состояние



Мониторинги - CD

Ansible-juggler

- › Читаемый yaml-file
- › Правила агрегации и время срабатывания проверок
- › Актуализирует состояние
- › Выкладывает изменения
- › Подавляет всплески при выкладке



Сбор численных метрик

Apache metrics

Система хранения

- › Ganglia
- › Graphite sink
- › Jmx2graphite

Визуализация

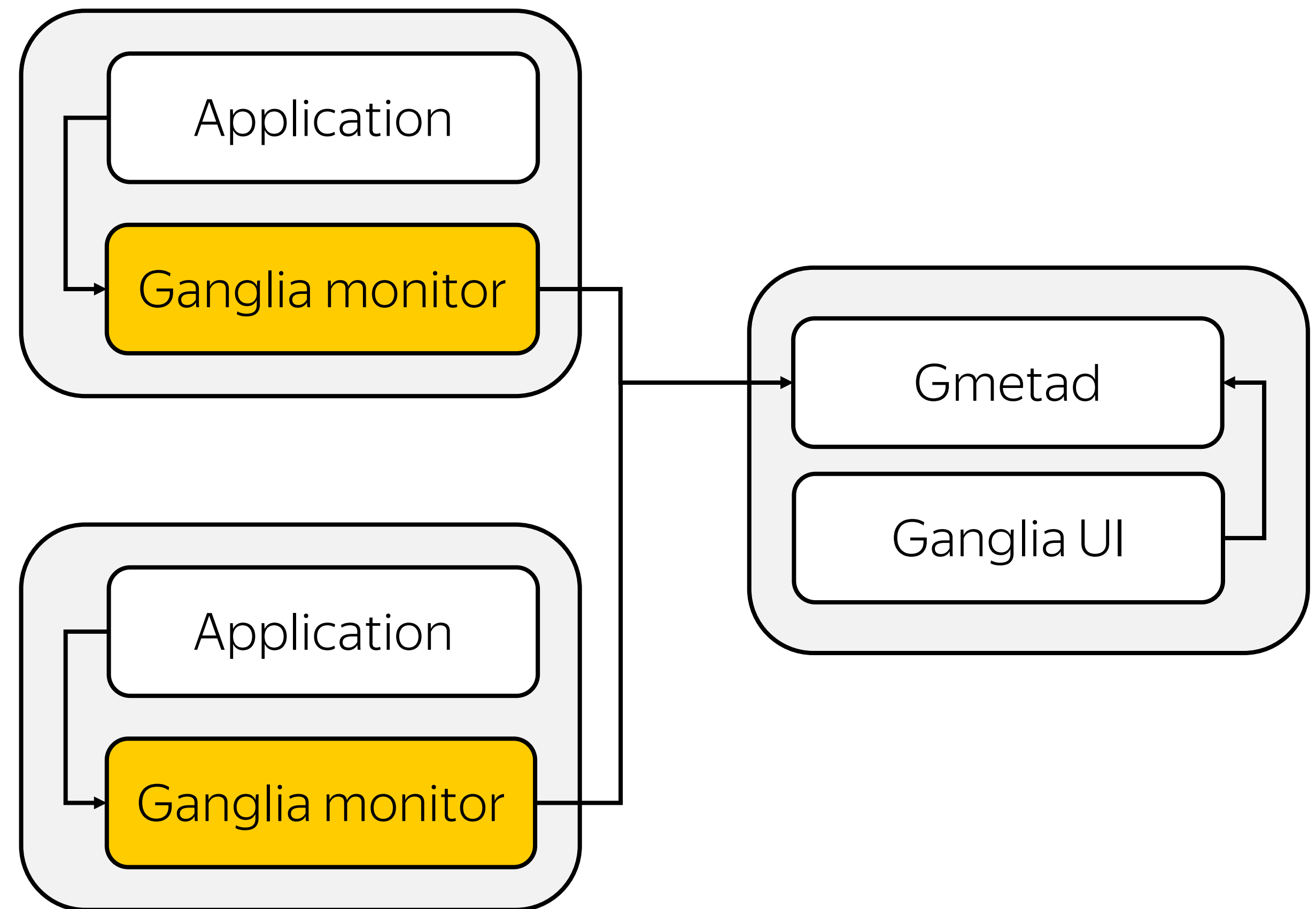
Сбор численных метрик

Apache metrics

Система хранения

- › Ganglia
- › Graphite sink
- › Jmx2graphite

Визуализация



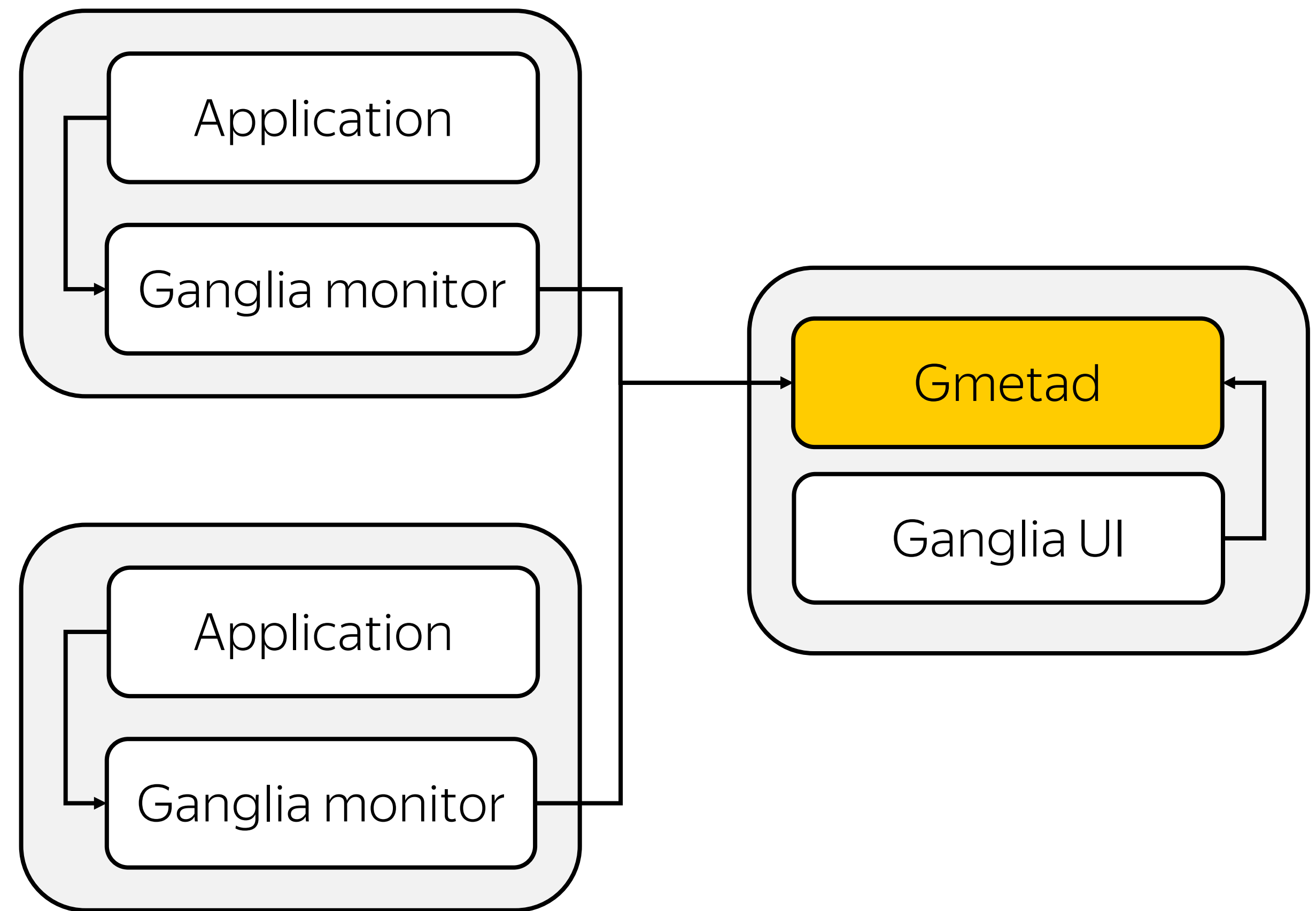
Сбор численных метрик

Apache metrics

Система хранения

- › Ganglia
- › Graphite sink
- › Jmx2graphite

Визуализация



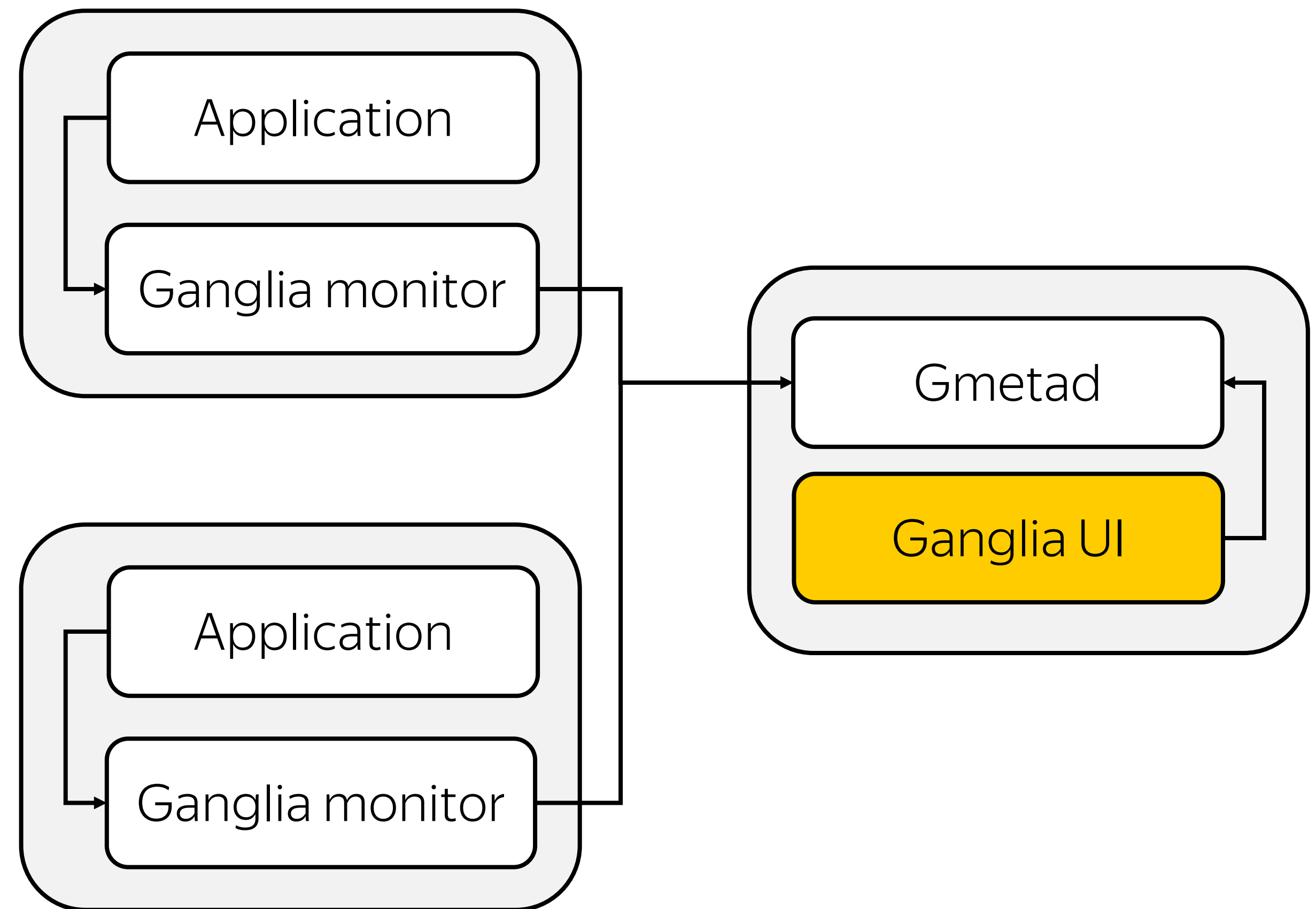
Сбор численных метрик

Apache metrics

Система хранения

- › Ganglia
- › Graphite sink
- › Jmx2graphite

Визуализация



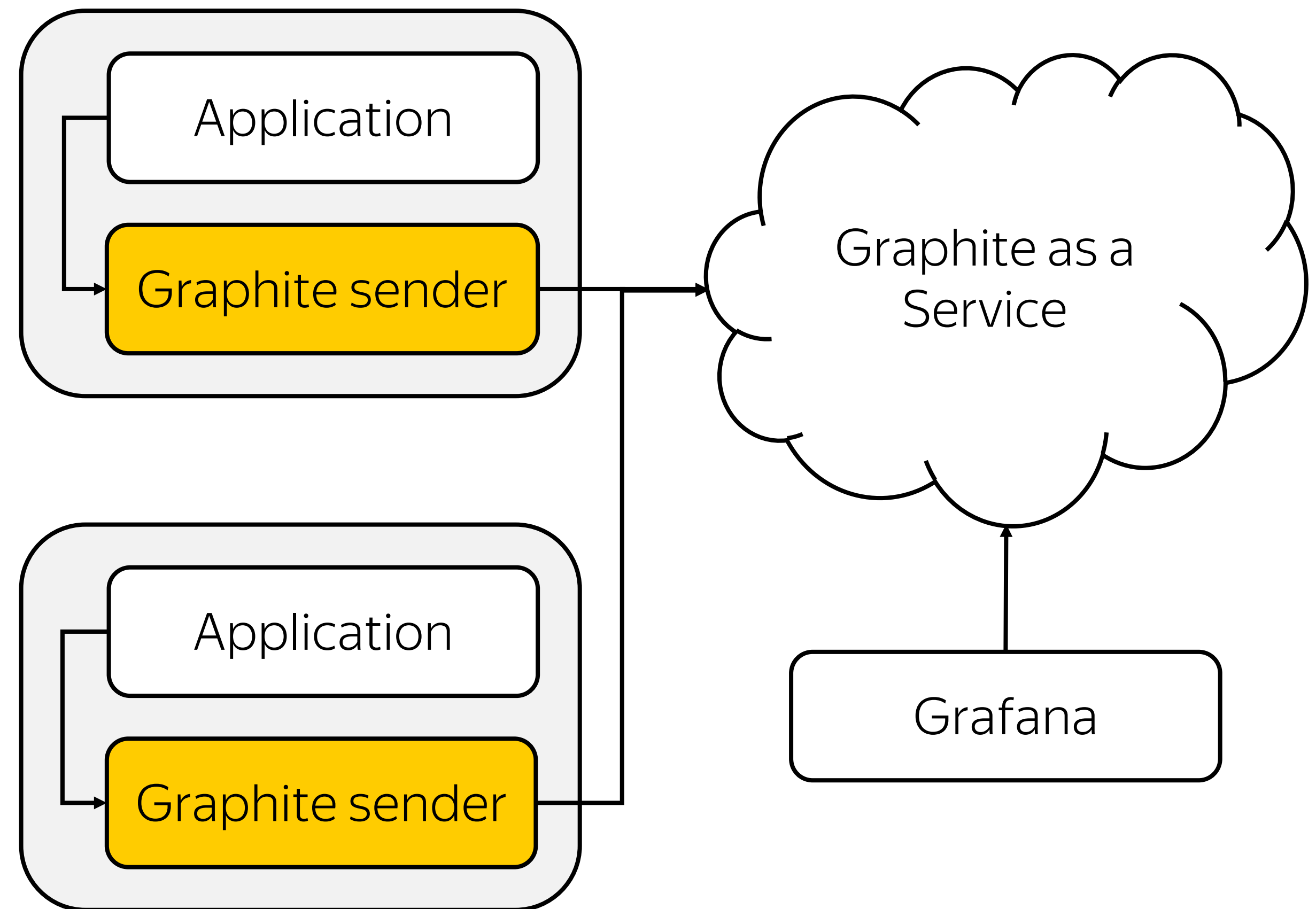
Сбор численных метрик

Apache metrics

Система хранения

- › Ganglia
- › Graphite sink
- › Jmx2graphite

Визуализация



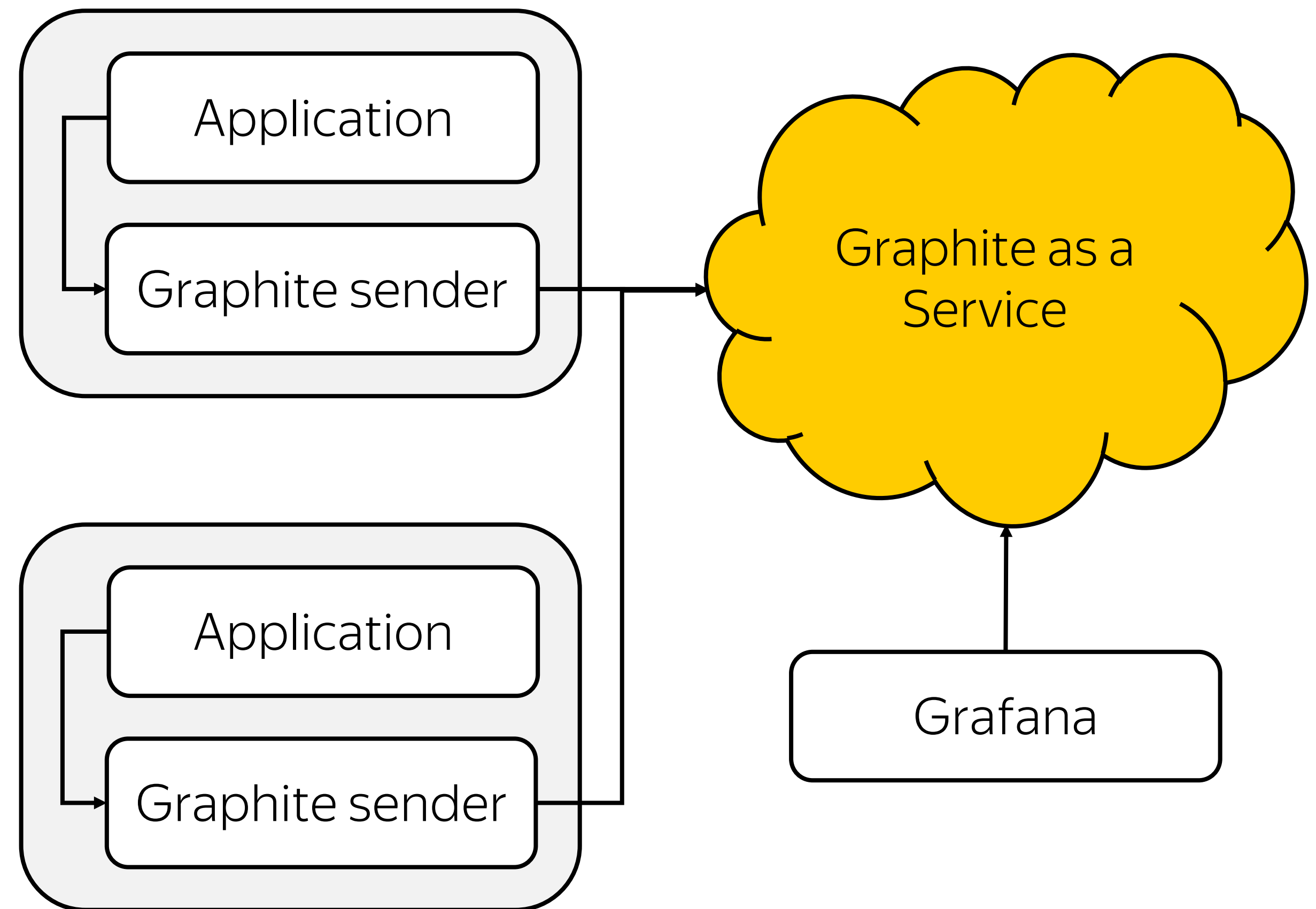
Сбор численных метрик

Apache metrics

Система хранения

- › Ganglia
- › Graphite sink
- › Jmx2graphite

Визуализация



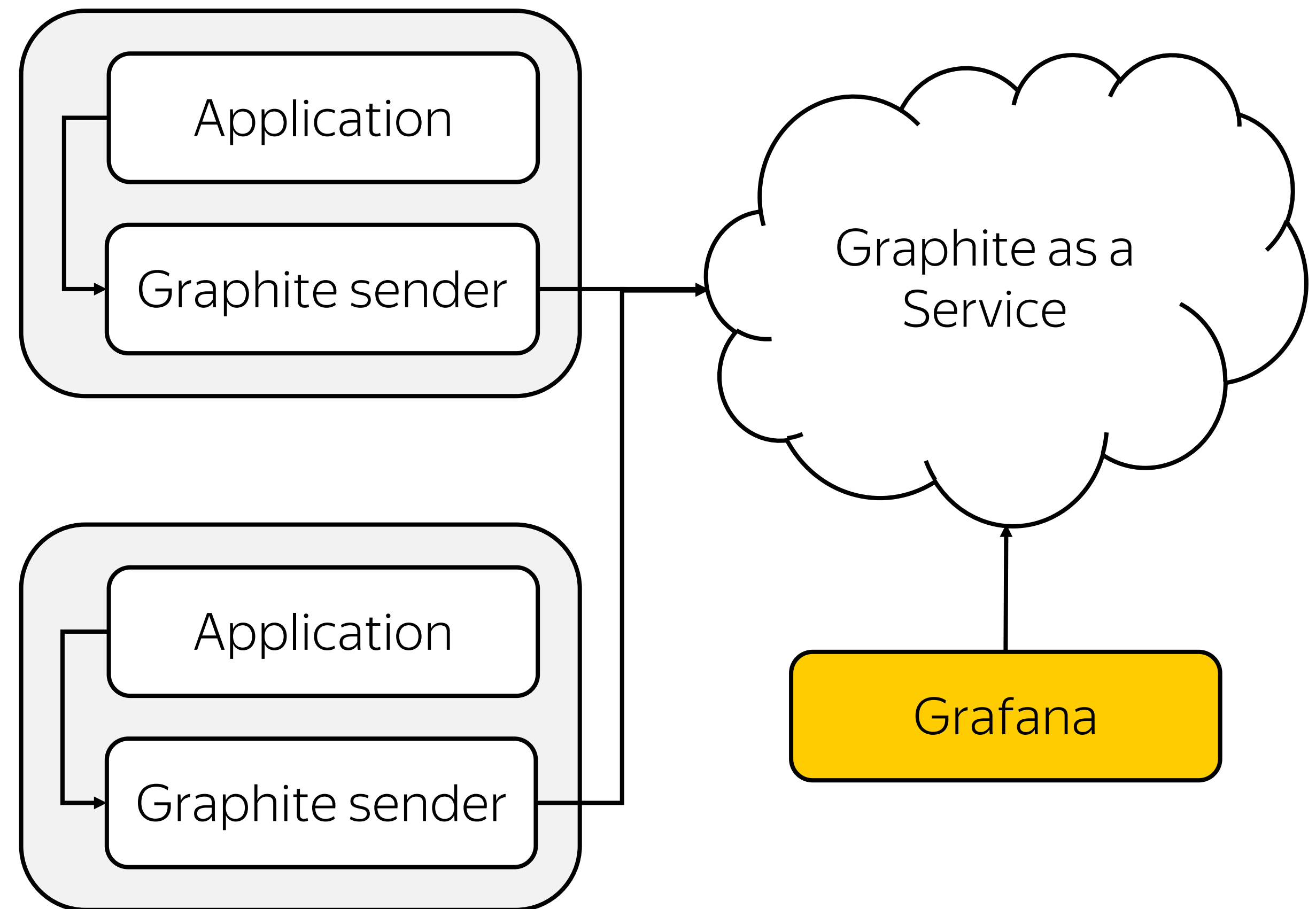
Сбор численных метрик

Apache metrics

Система хранения

- › Ganglia
- › Graphite sink
- › Jmx2graphite

Визуализация



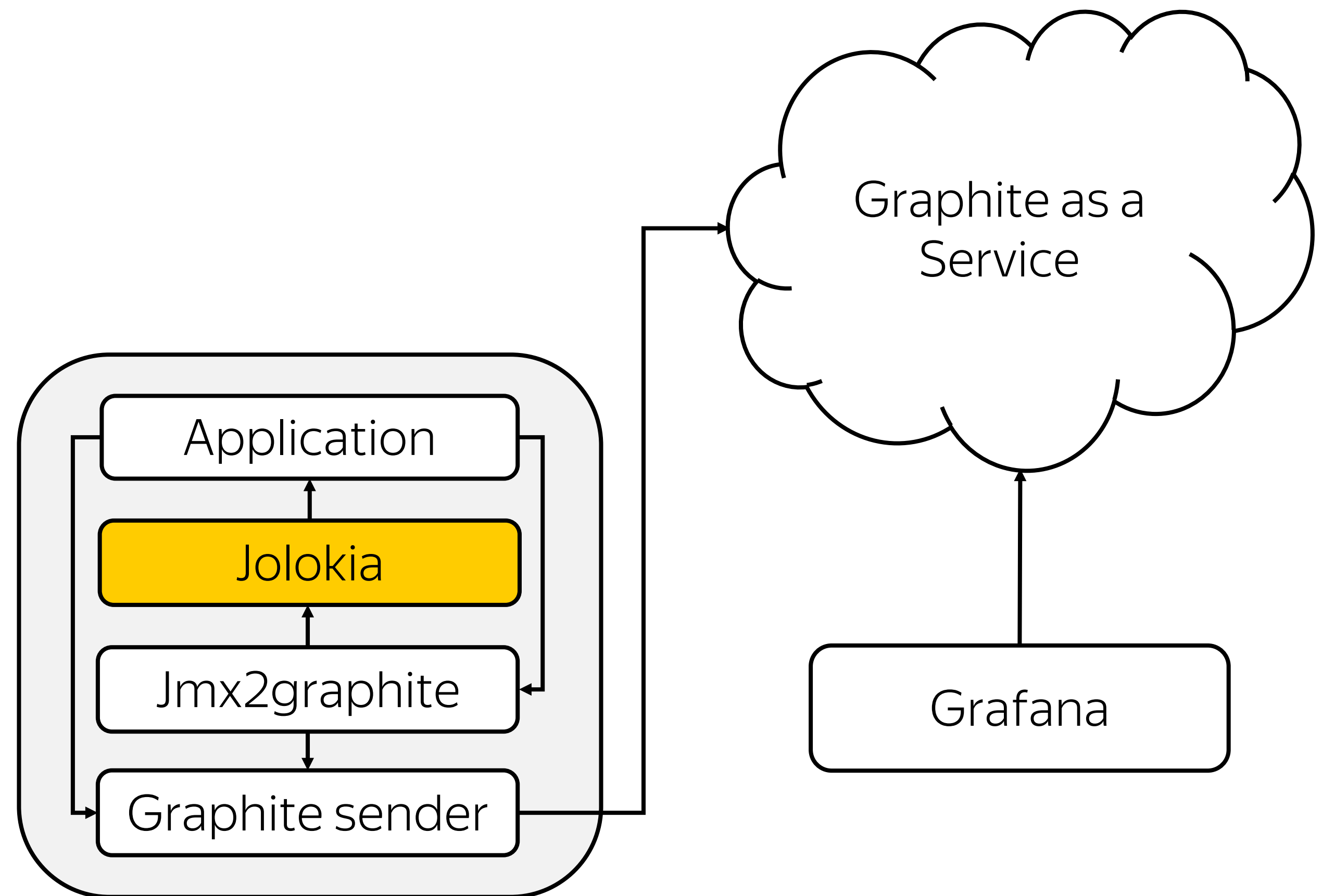
Сбор численных метрик

Apache metrics

Система хранения

- › Ganglia
- › Graphite sink
- › Jmx2graphite

Визуализация



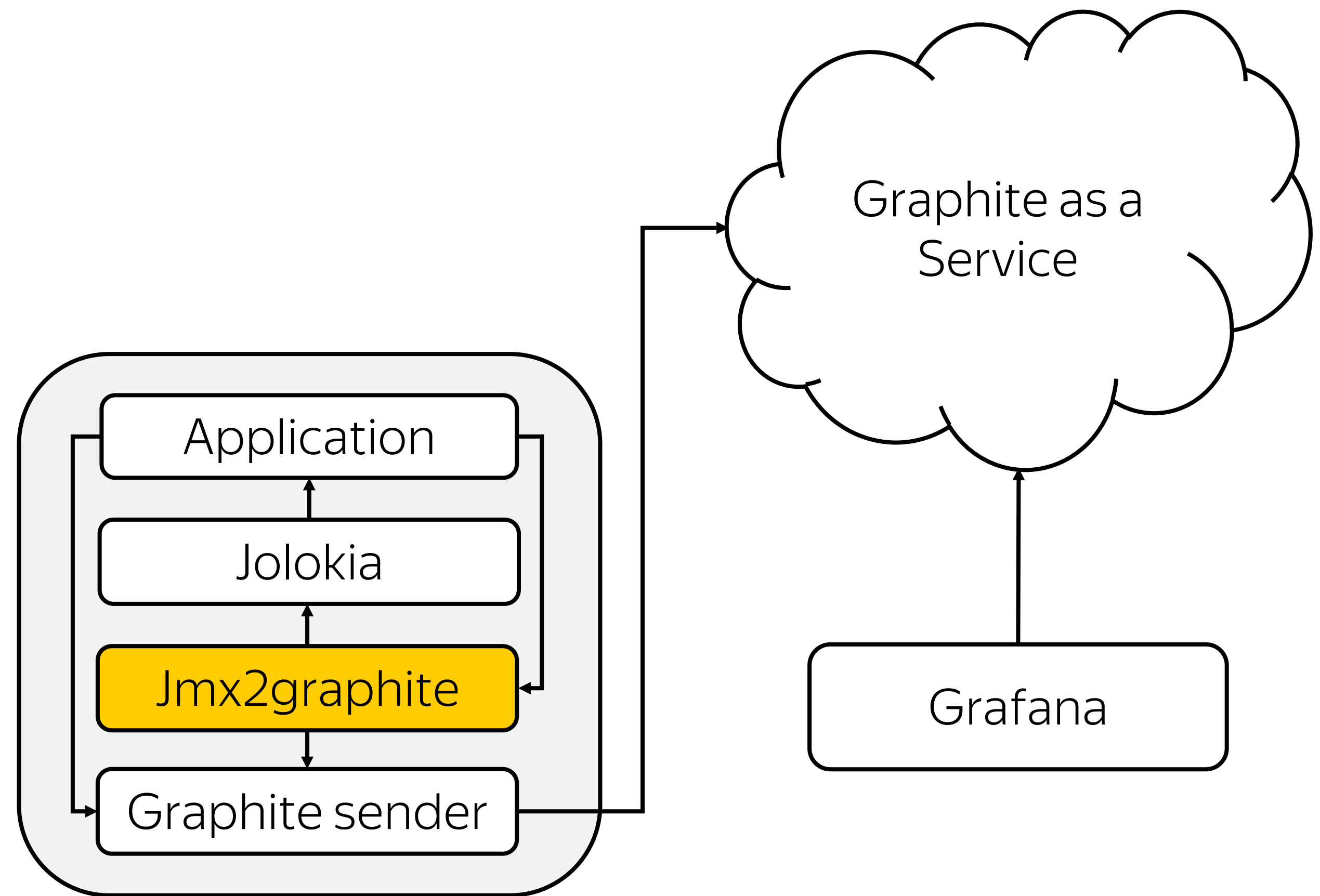
Сбор численных метрик

Apache metrics

Система хранения

- › Ganglia
- › Graphite sink
- › Jmx2graphite

Визуализация



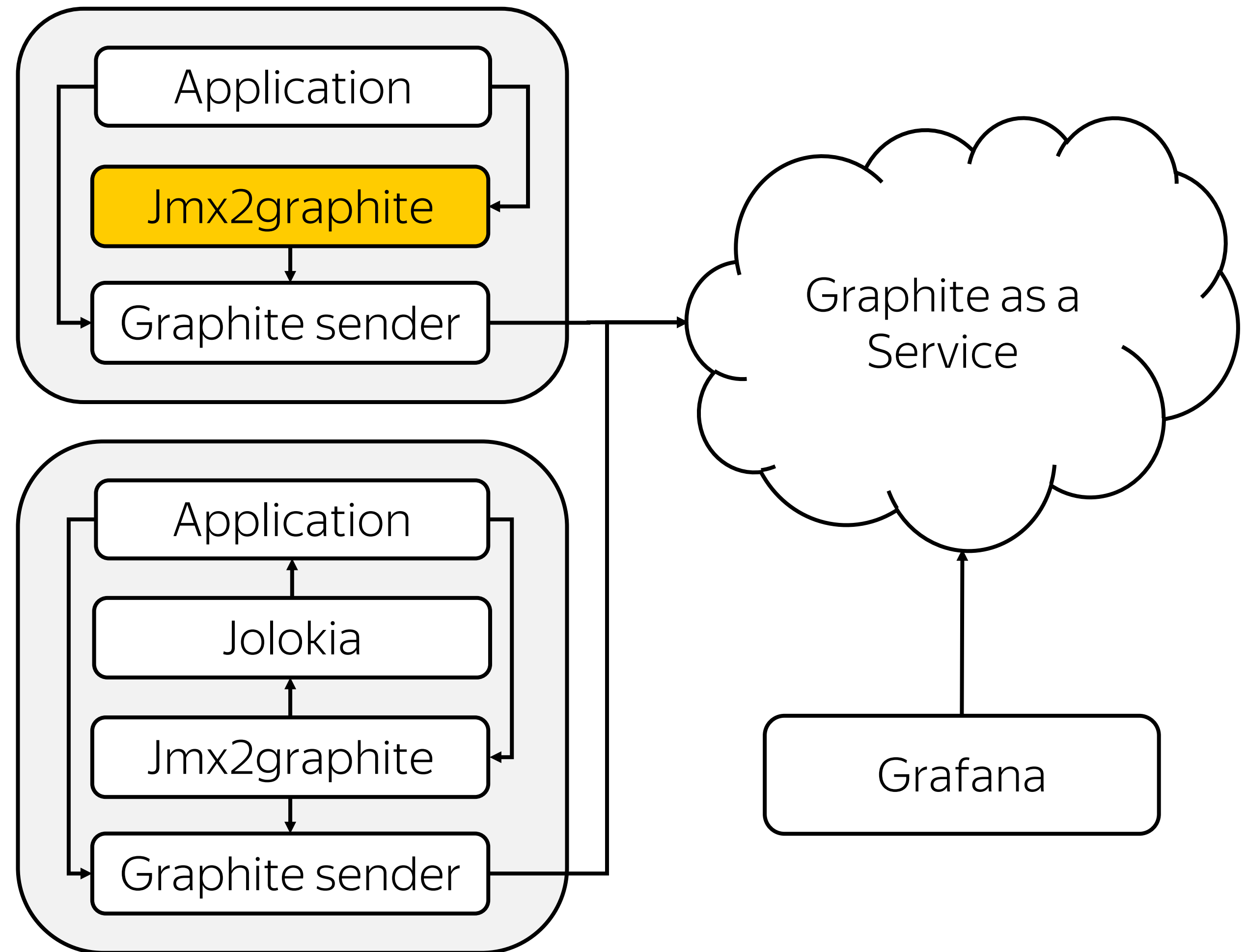
Сбор численных метрик

Apache metrics

Система хранения

- › Ganglia
- › Graphite sink
- › Jmx2graphite

Визуализация

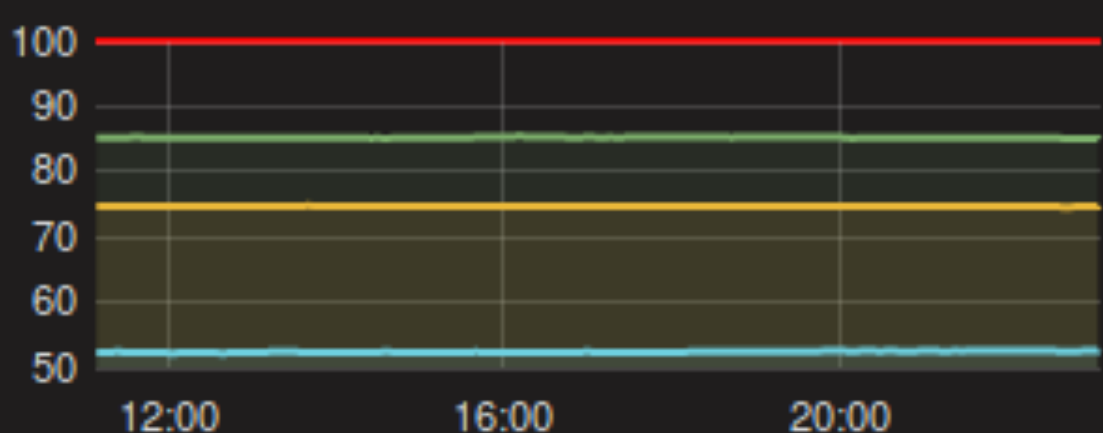


Дашборды

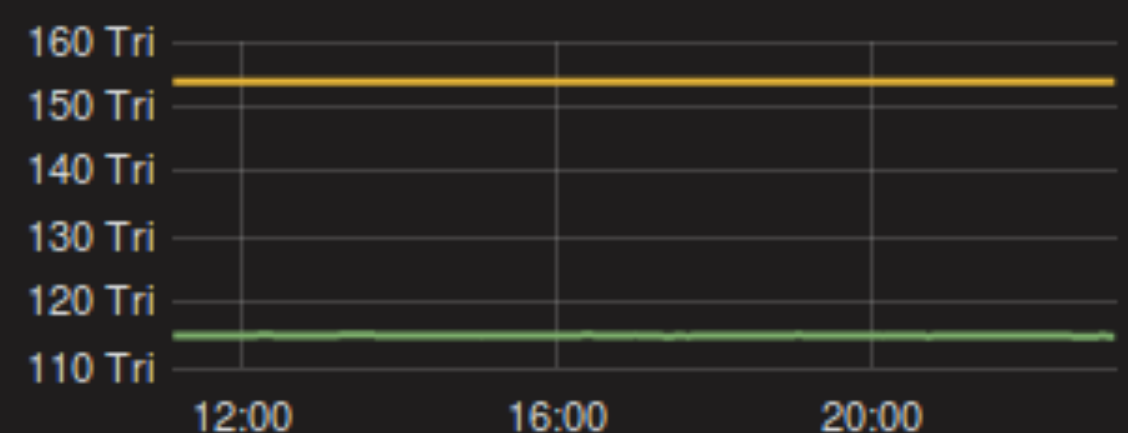
- Выделенный дашборд на каждый кластер
- Общий дашборд с KPI заказчиков
- Специализированные дашборды для анализа подсистем



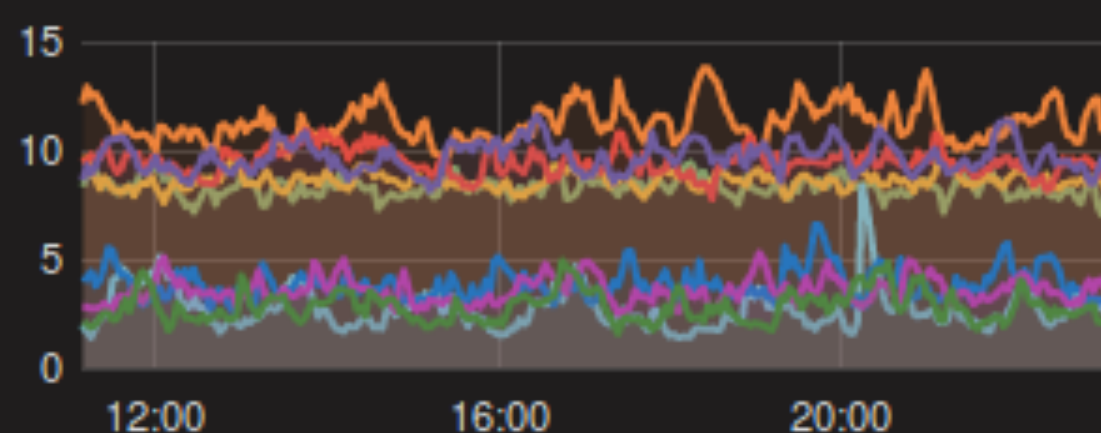
diskspace (percent)



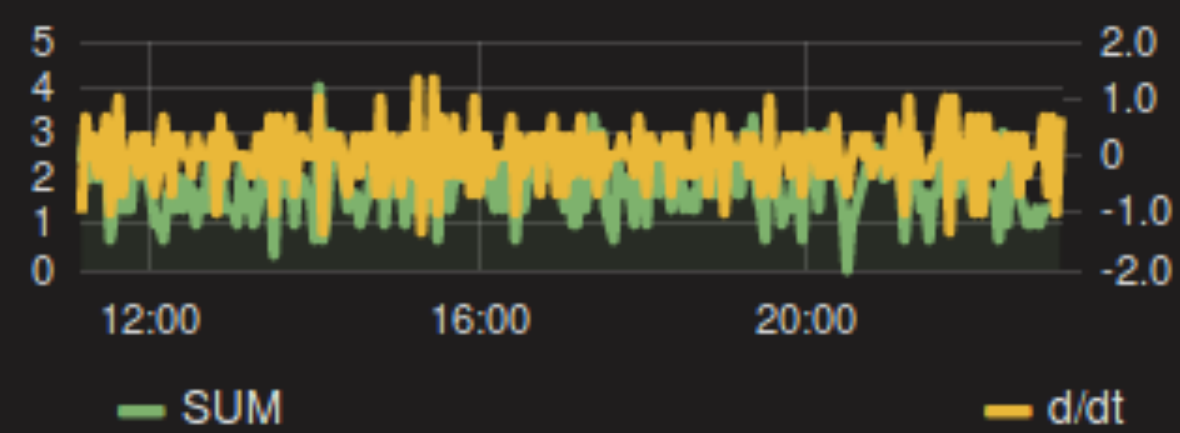
diskspace (Tb)



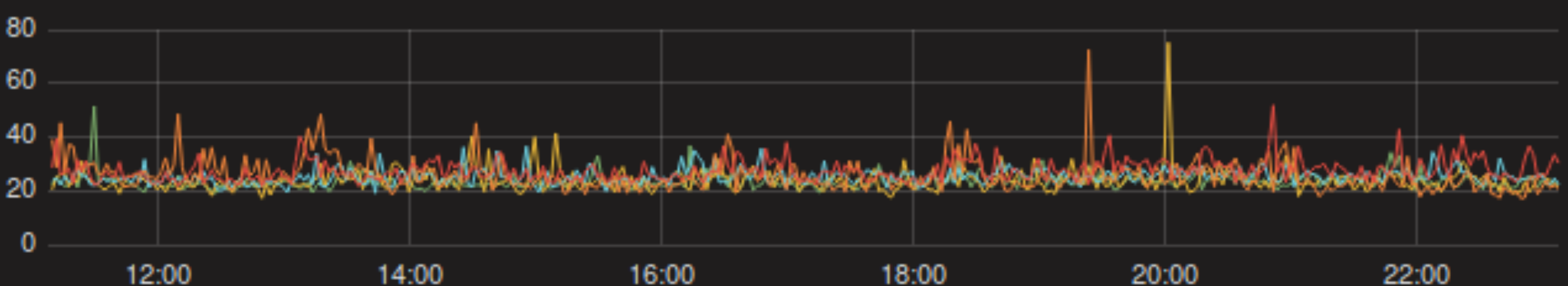
la



opentsdb queueSize



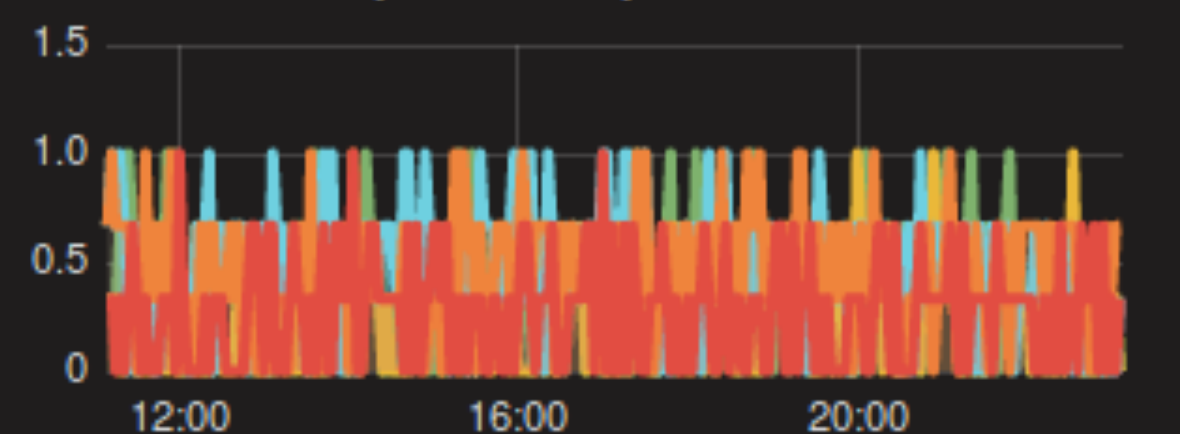
diskutil



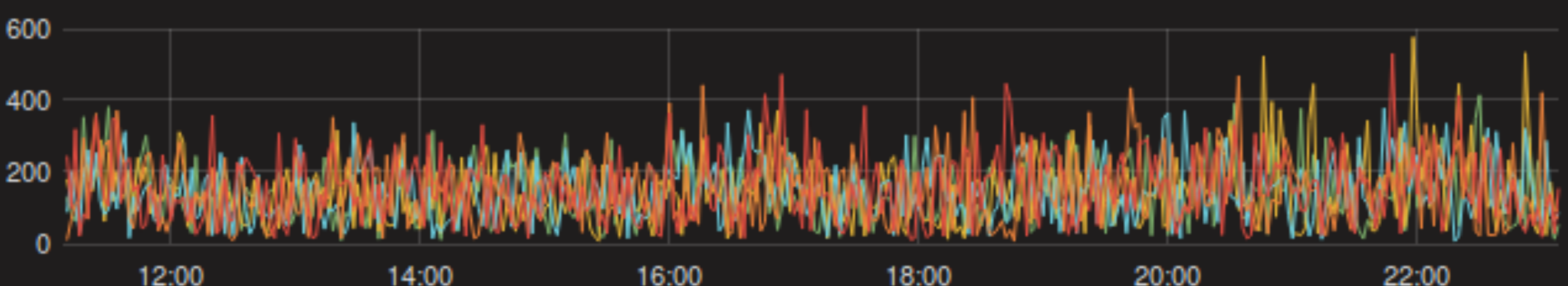
cpu (average)



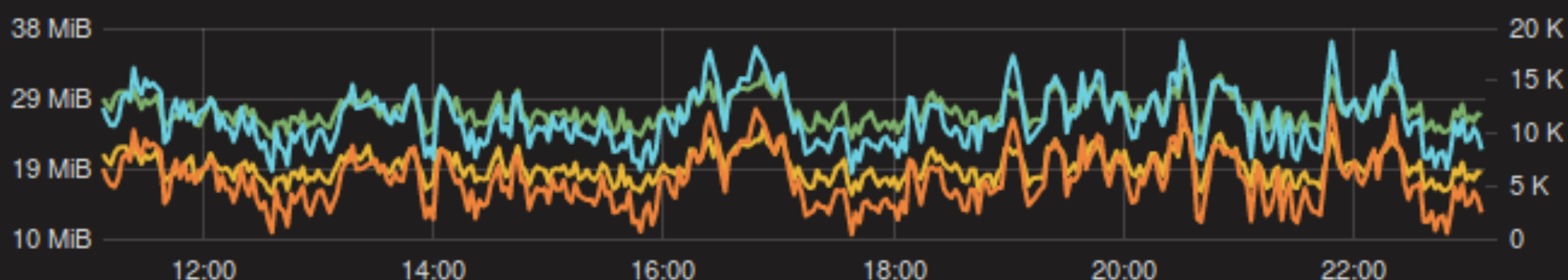
opentsdb queueSize



diskwait



network bytes packets (average)



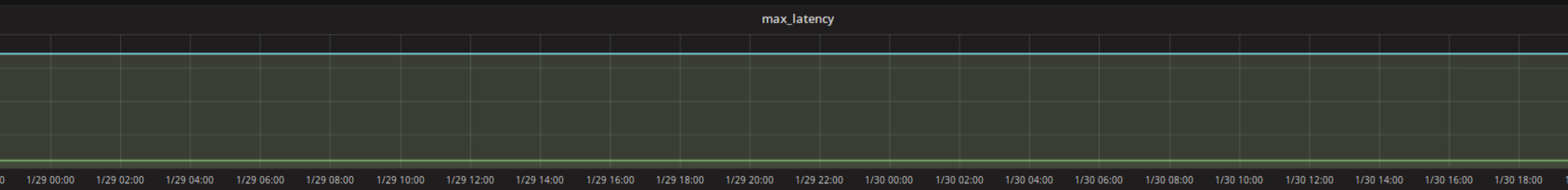
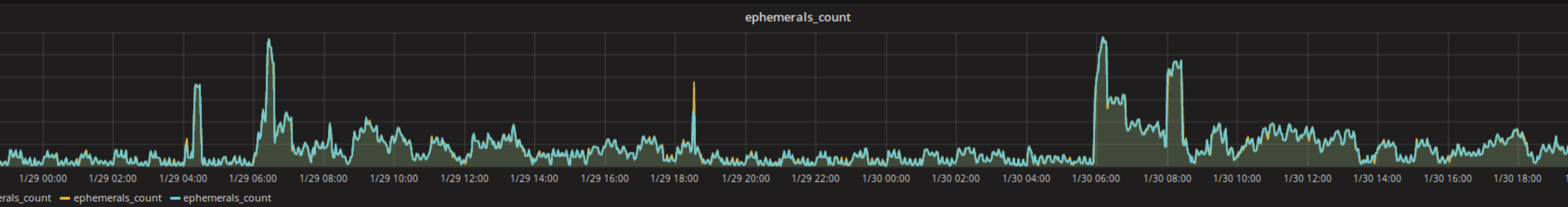
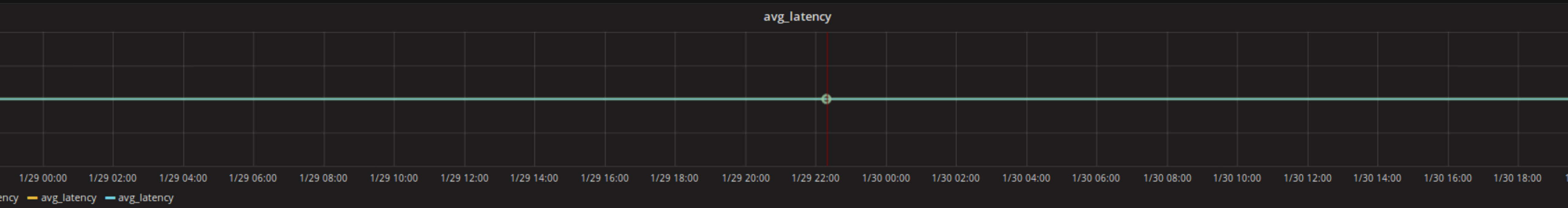
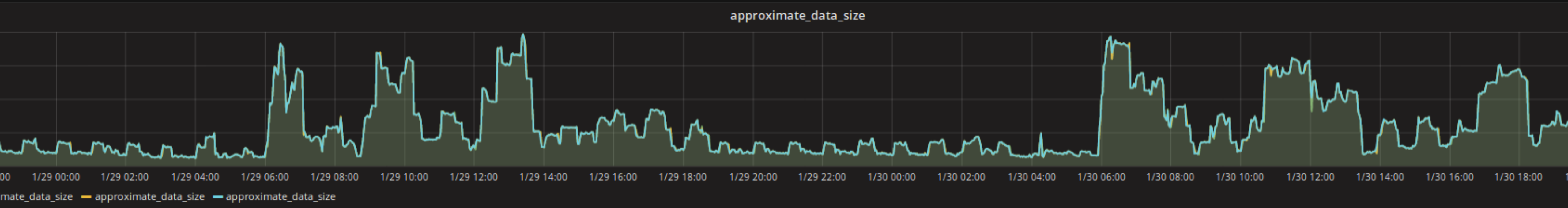
+ ADD ROW

Дашборды

- Выделенный дашборд на каждый кластер
- Общий дашборд с KPI заказчиков
- Специализированные дашборды для анализа подсистем

Дашборды

- Выделенный дашборд на каждый кластер
- Общий дашборд с KPI заказчиков
- Специализированные дашборды для анализа подсистем



Проблемы и решения



Проблема N°1 – управление и синхронизация

Проблема №1 – управление и синхронизация

- Выкладка изменений на кластер

- Переиспользование кода

- Git merge hell

Проблема №1 – управление и синхронизация

Выкладка изменений на кластер

rsync

Переиспользование кода

Отдельный репозиторий на кластер
Общие submodules

Git merge hell

Git flow

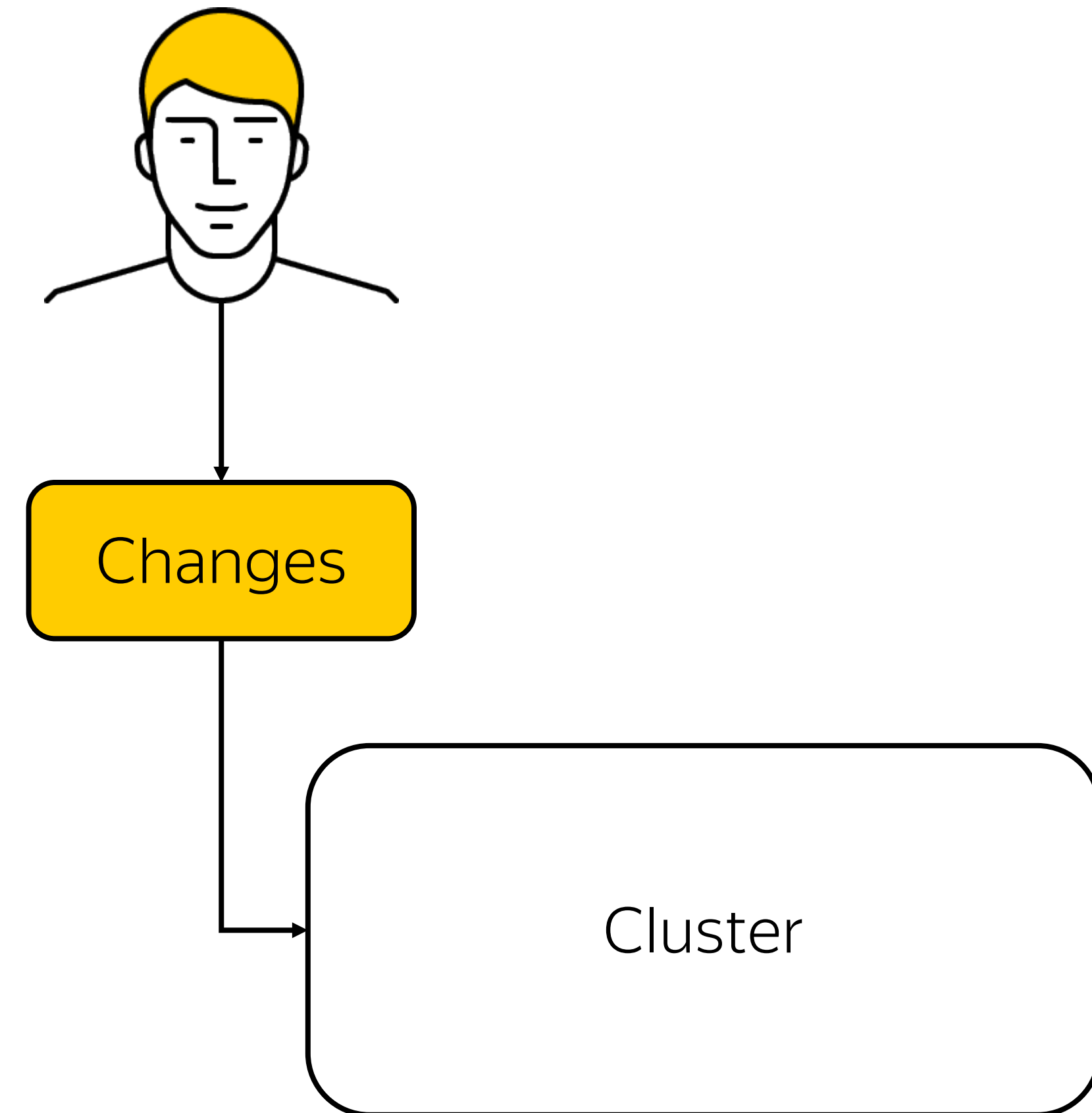
Проблема №1 – управление и синхронизация

Выкладка изменений на кластер

Переиспользование кода

Git merge hell

Синхронизация между администраторами



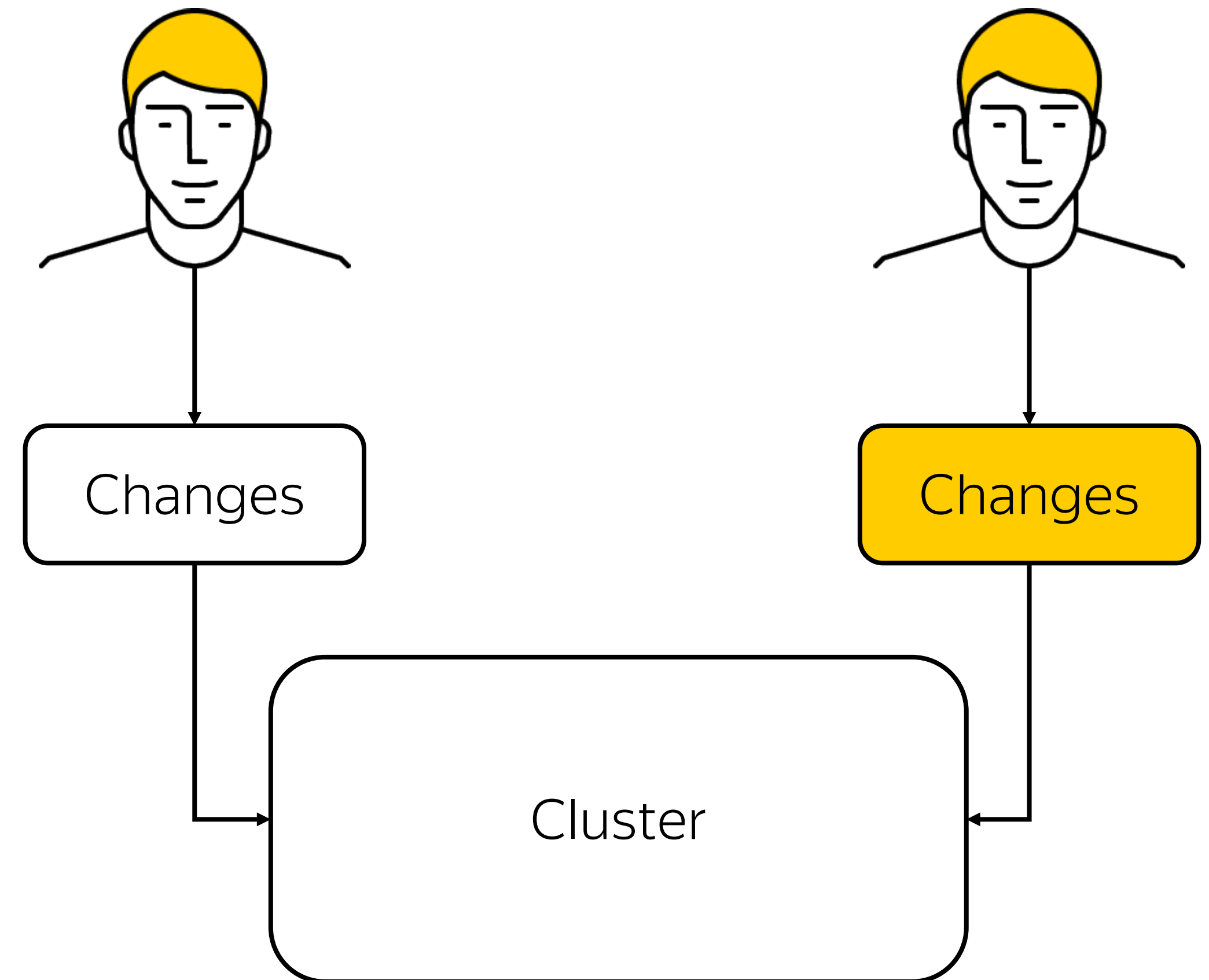
Проблема №1 – управление и синхронизация

Выкладка изменений на кластер

Переиспользование кода

Git merge hell

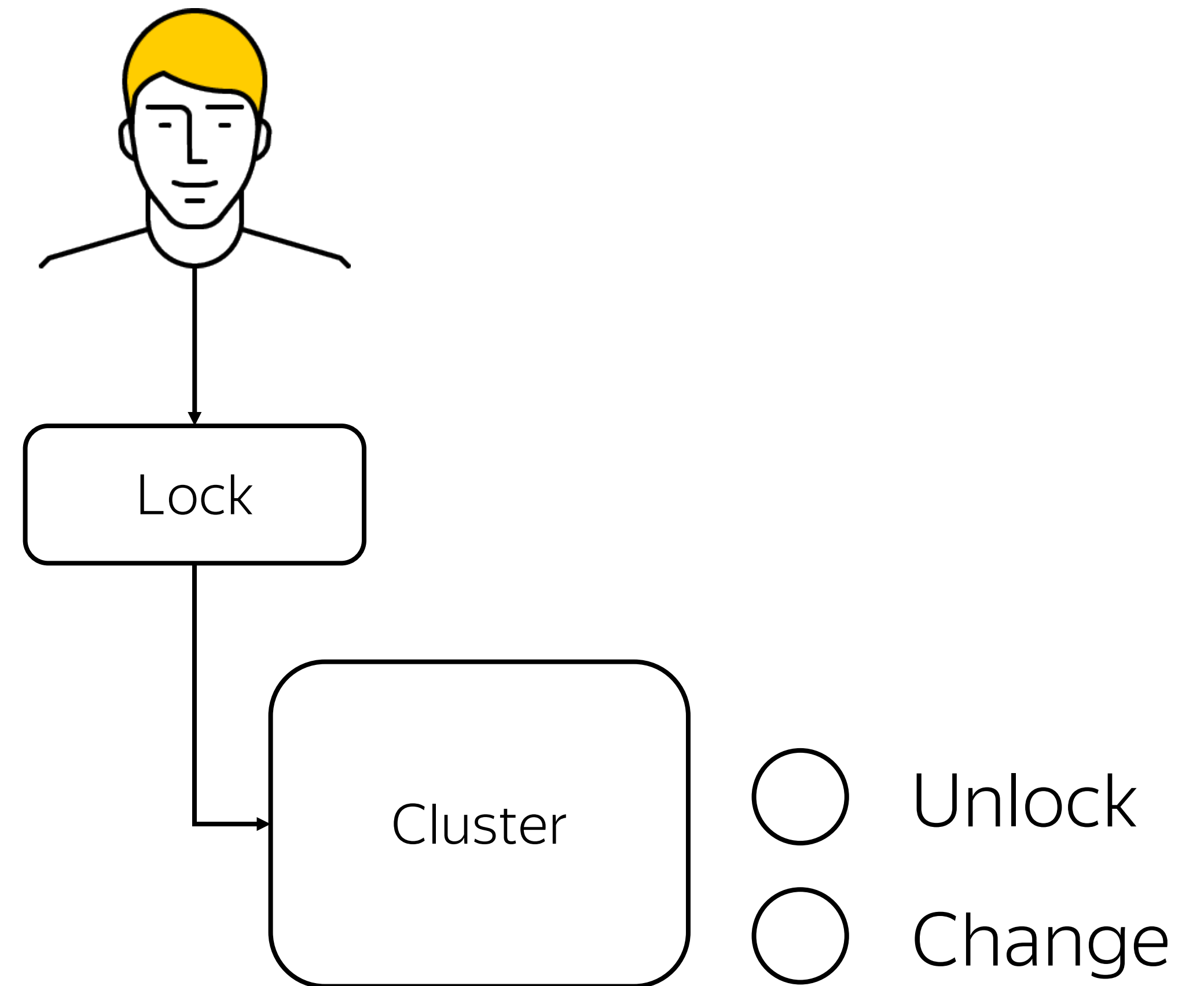
Синхронизация между администраторами



Salto

Salto:

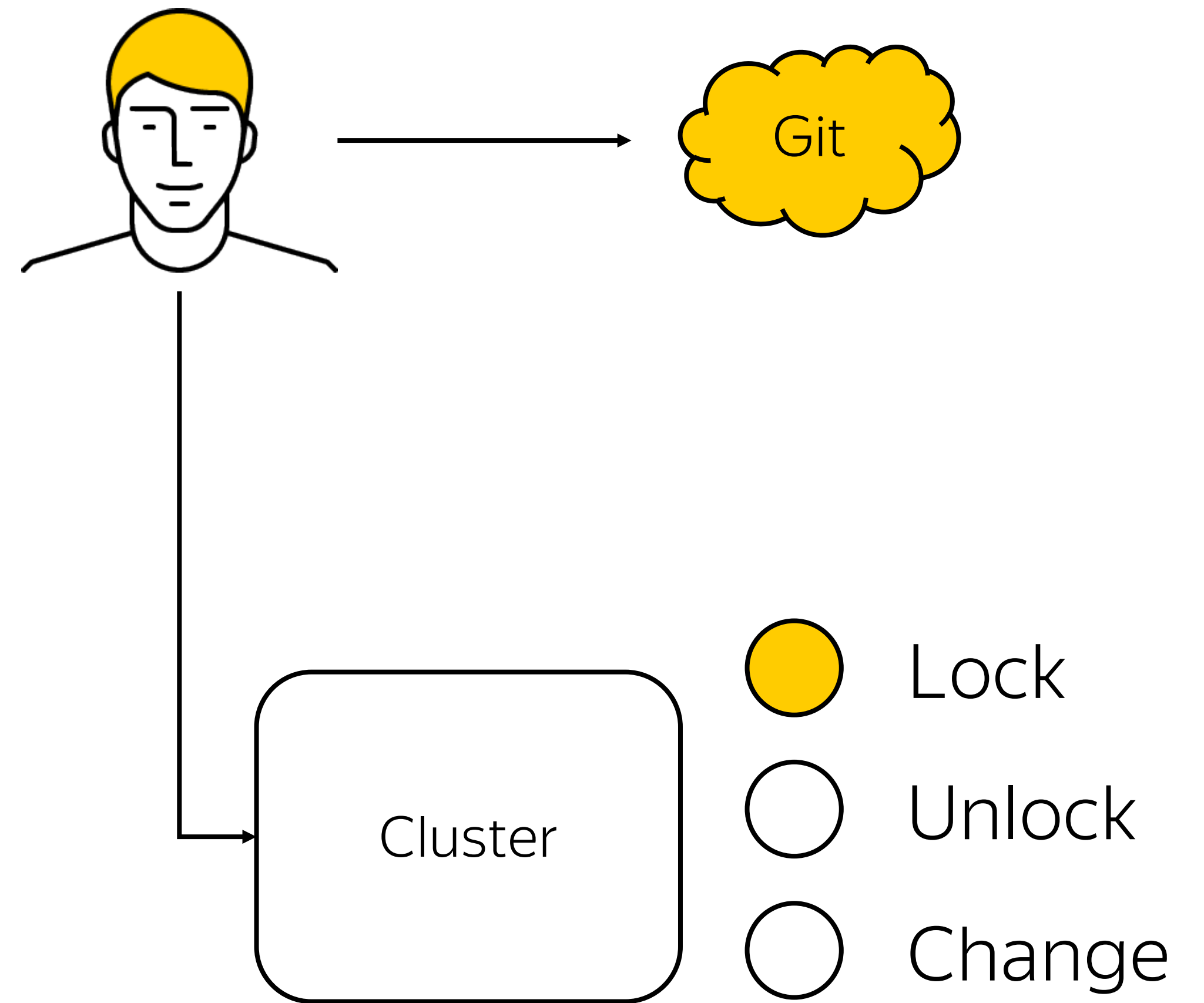
› Mutex поверх git



Salto

Salto:

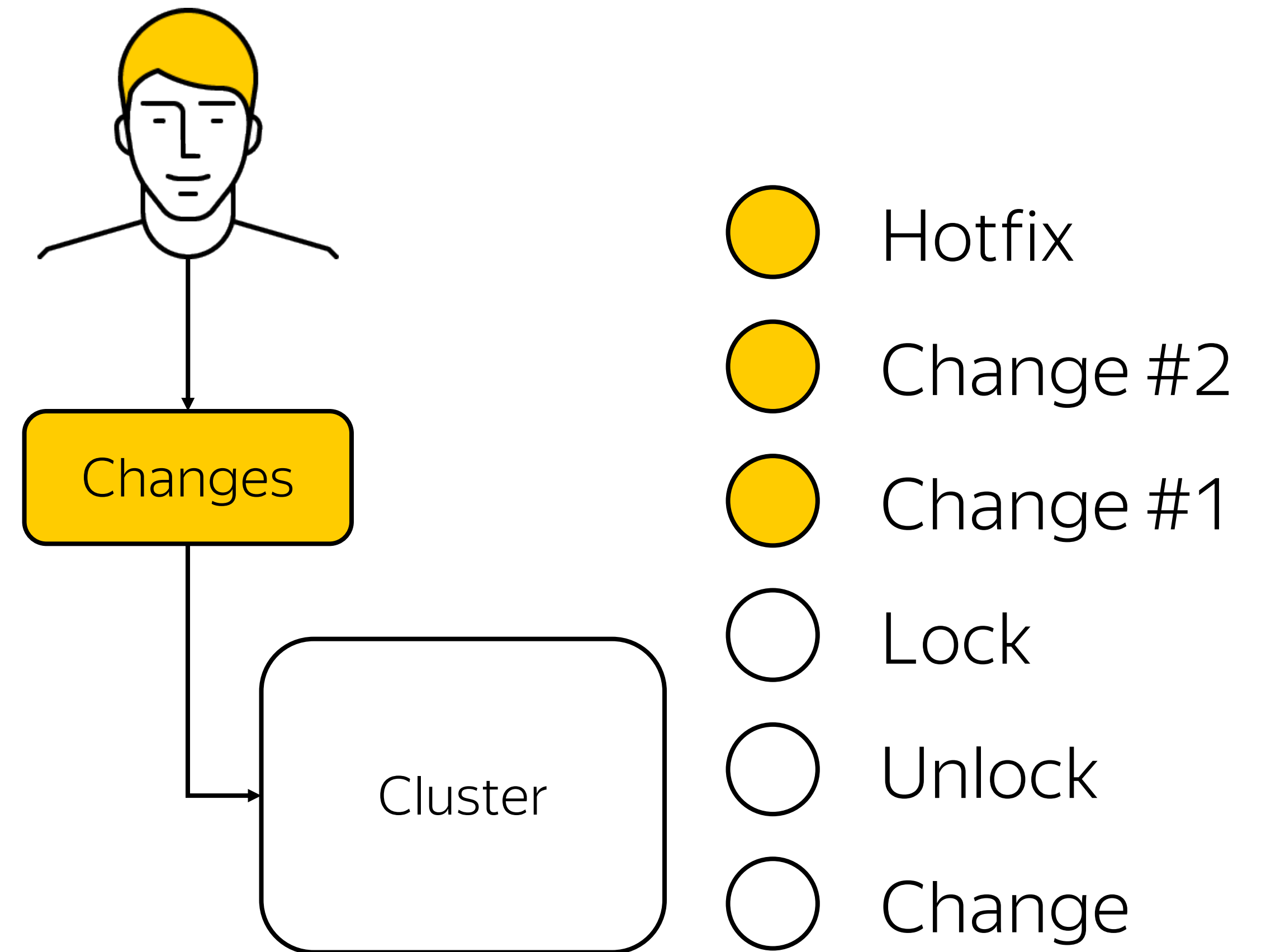
› Mutex поверх git



Salto

Salto:

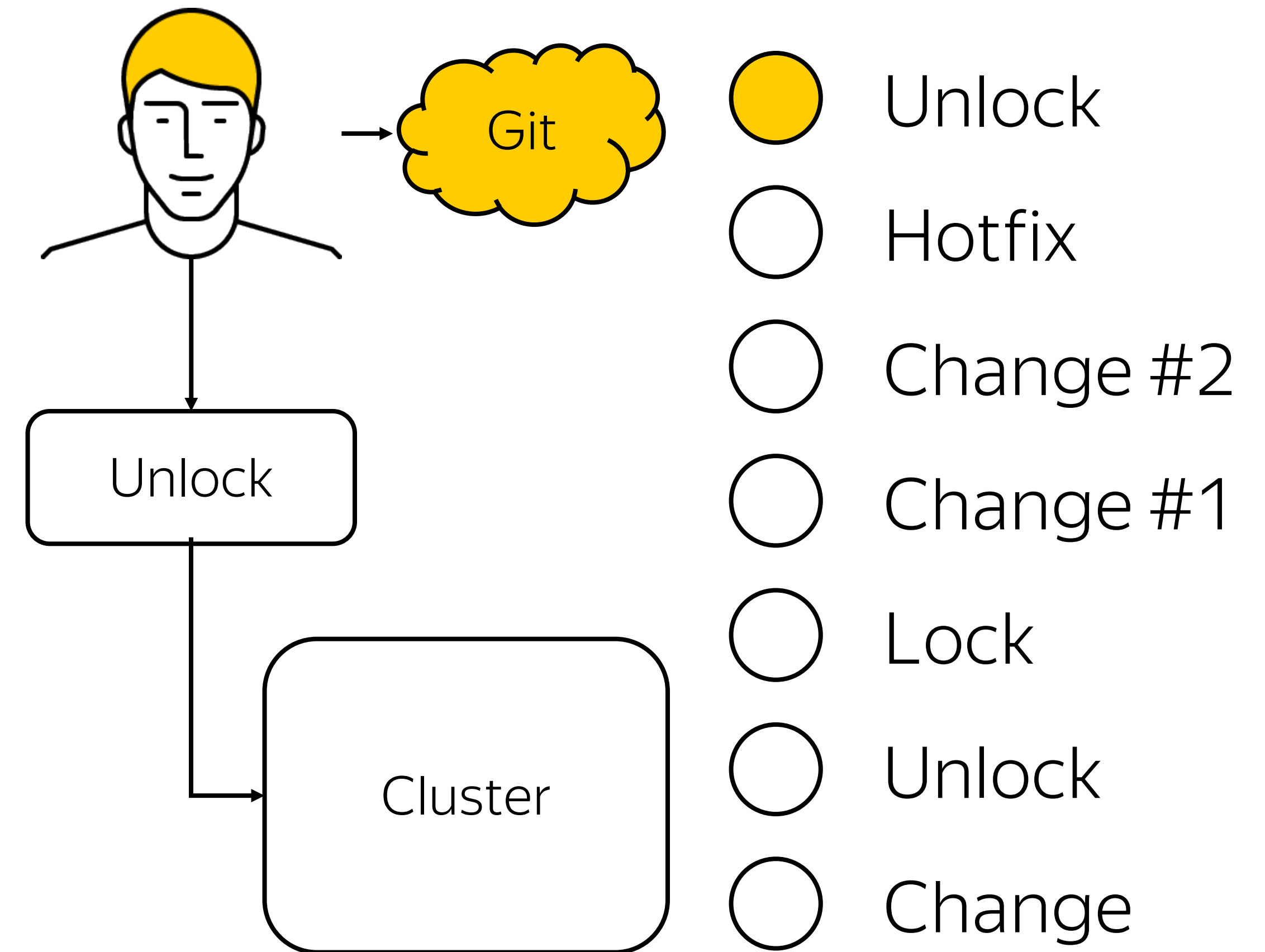
› Mutex поверх git



Salto

Salto:

- › Mutex поверх git
- › Обновление зависимостей до актуальной версии



Проблема 2: обновления и performance

Проблема 2: обновления и performance

Обновления:

- › На новую версию Hadoop
- › Между релизами
- › Хотфиксы

Зависимости в приложениях

Проблема 2: обновления и performance

Обновления:

- › На новую версию Hadoop
- › Между релизами
- › Хотфиксы

Зависимости в приложениях

Настройки у пользователей

Проблема 2: обновления и performance

Обновления:

- › На новую версию Hadoop
- › Между релизами
- › Хотфиксы

Зависимости в приложениях

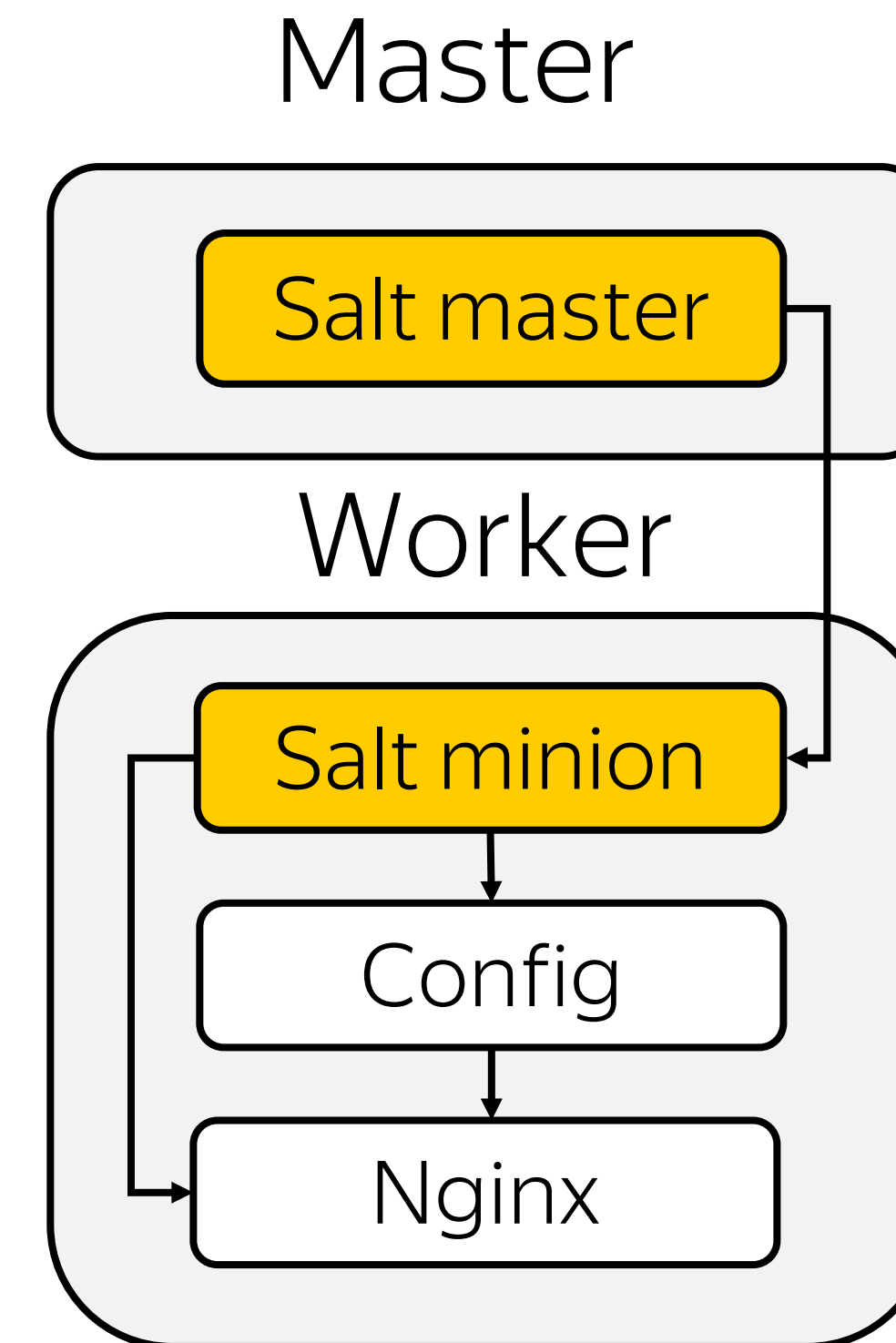
Настройки у пользователей

Производительность

Обновление конфигурации для пользователей

CI со сборкой пакетов

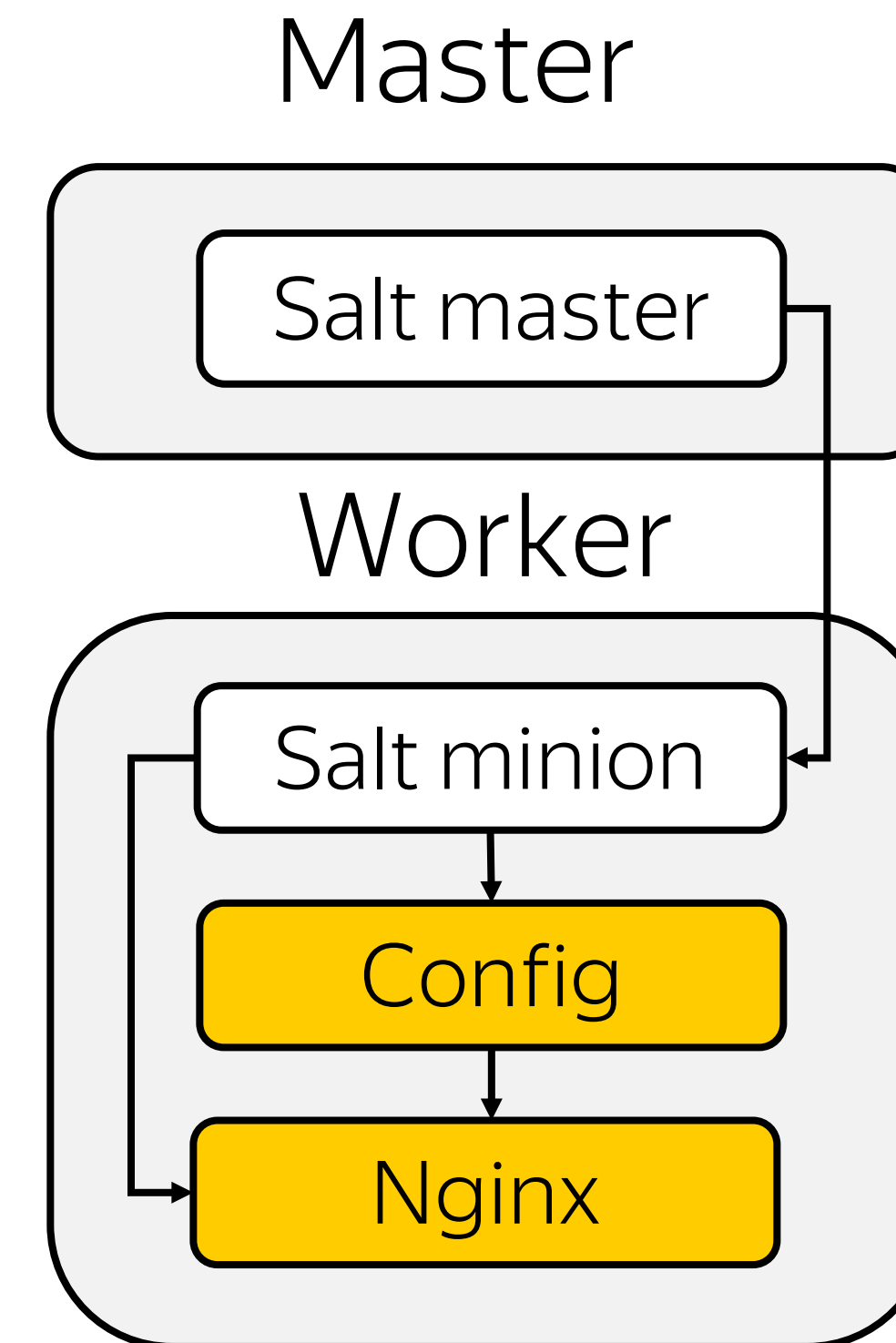
- › Рендер конфигурации из salt
- › Nginx с архивом конфигурации
- › Обновление по state.highstate



Обновление конфигурации для пользователей

CI со сборкой пакетов

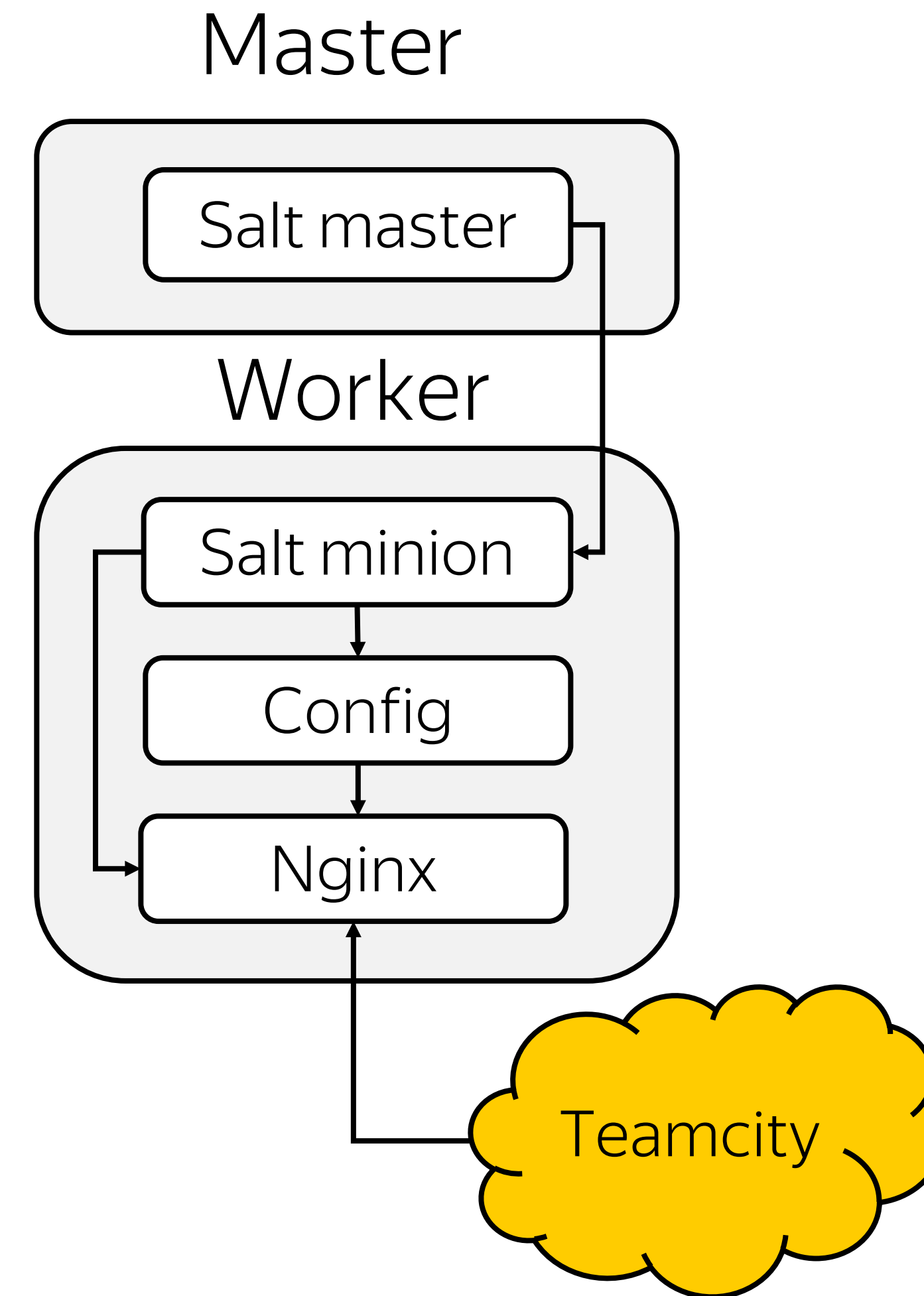
- › Рендер конфигурации из salt
- › Nginx с архивом конфигурации
- › Обновление по state.highstate



Обновление конфигурации для пользователей

CI со сборкой пакетов

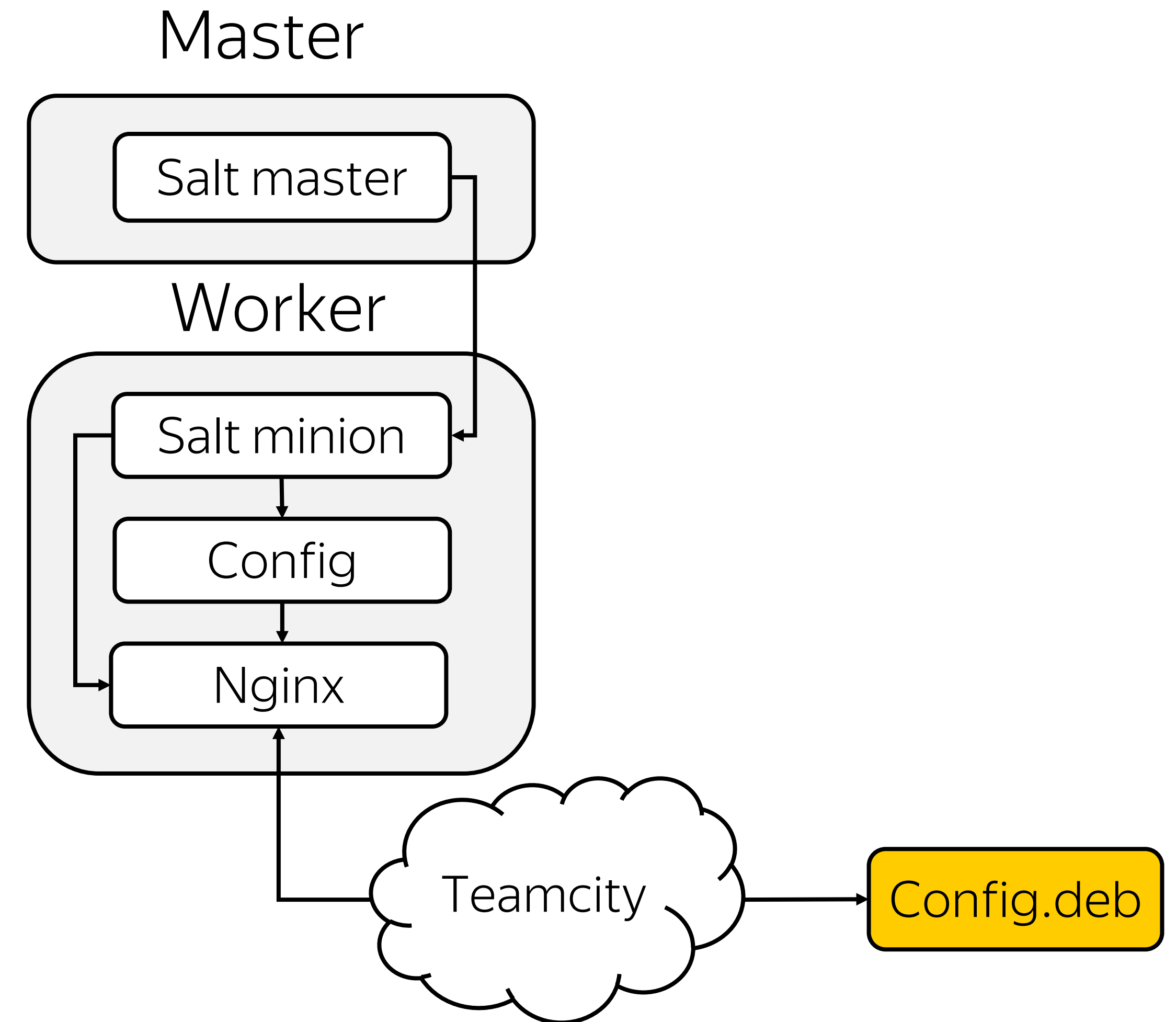
- › Рендер конфигурации из salt
- › Nginx с архивом конфигурации
- › Обновление по state.highstate



Обновление конфигурации для пользователей

CI со сборкой пакетов

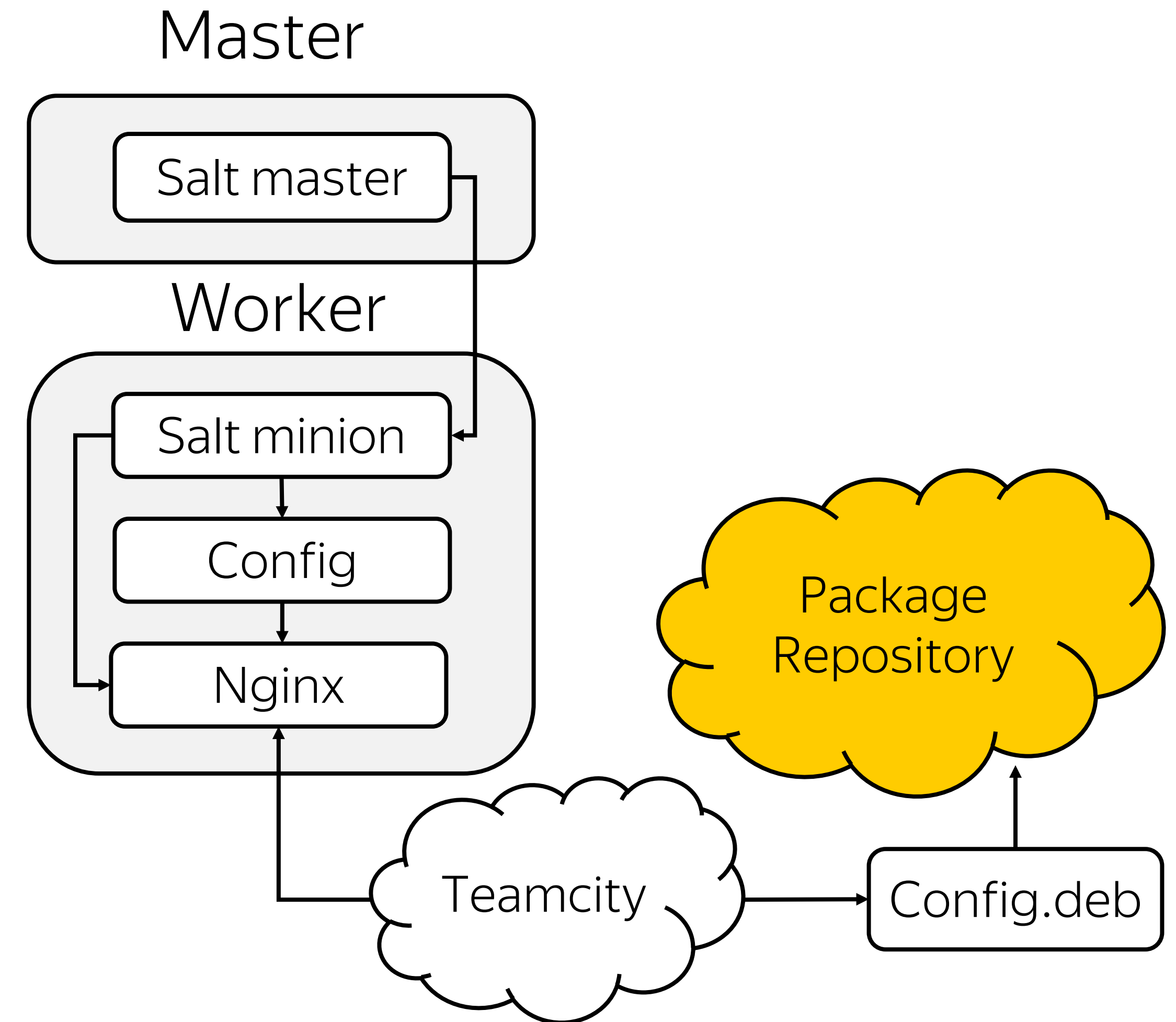
- › Рендер конфигурации из salt
- › Nginx с архивом конфигурации
- › Обновление по state.highstate



Обновление конфигурации для пользователей

CI со сборкой пакетов

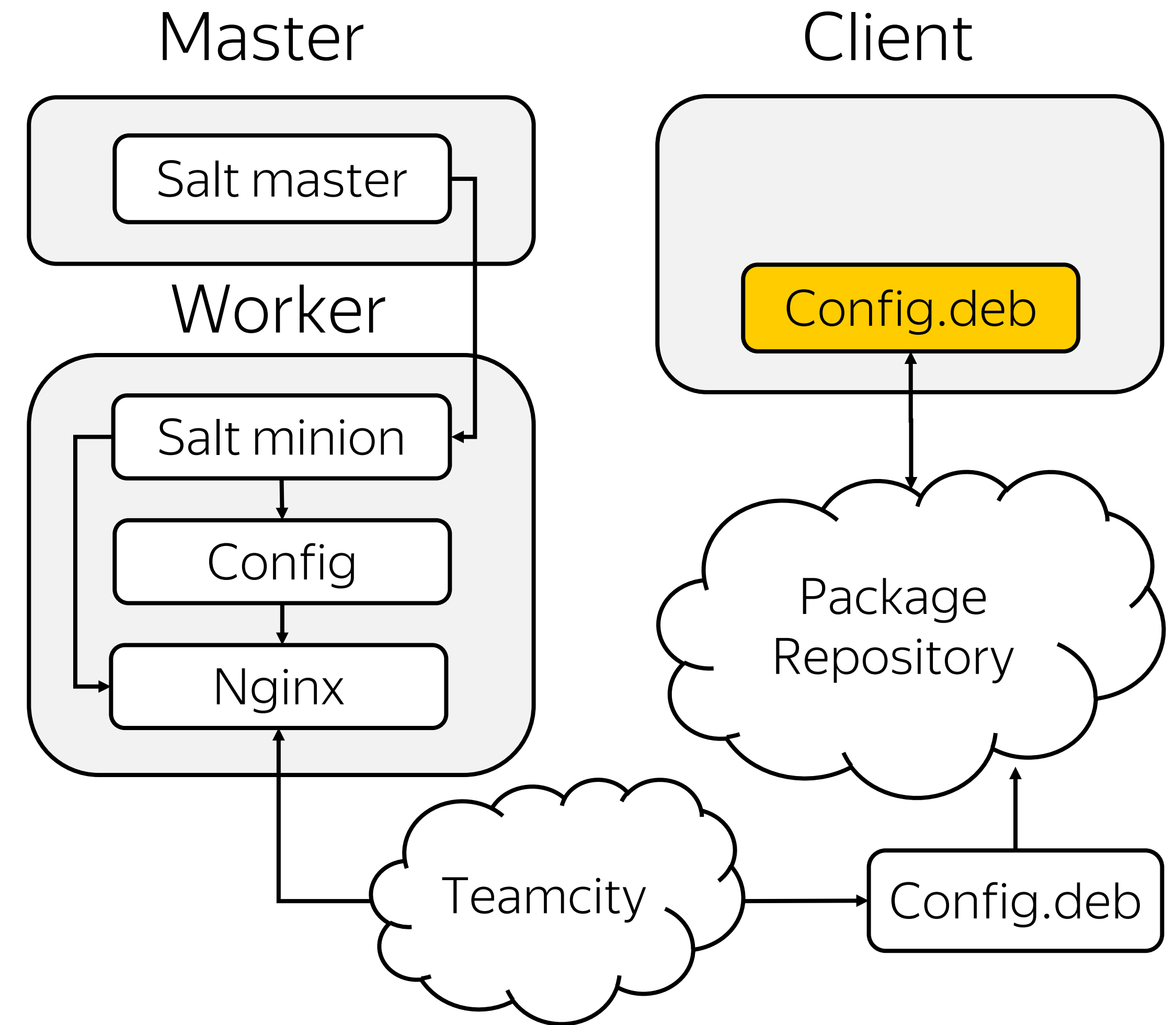
- › Рендер конфигурации из salt
- › Nginx с архивом конфигурации
- › Обновление по state.highstate



Обновление конфигурации для пользователей

CI со сборкой пакетов

- › Рендер конфигурации из salt
- › Nginx с архивом конфигурации
- › Обновление по state.highstate

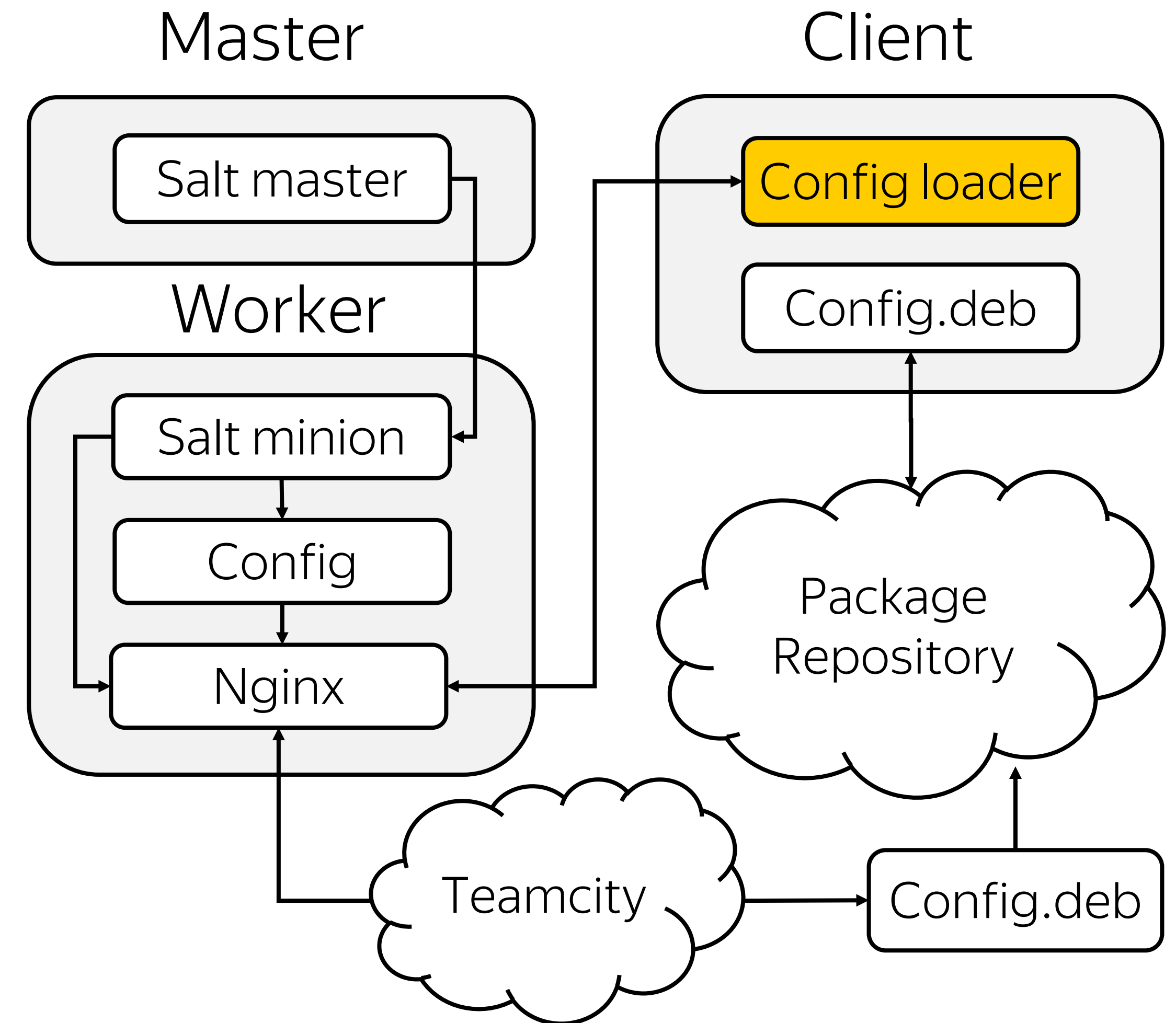


Обновление конфигурации для пользователей

CI со сборкой пакетов

- › Рендер конфигурации из salt
- › Nginx с архивом конфигурации
- › Обновление по state.highstate

CD с машин кластера



Тестирование производительности

■ Невозможно протестировать всю платформу

Тестирование производительности

- Невозможно протестировать всю платформу

- Тесты в dev-окружении

- › Синтетика – dfsio, terasort
- › Тесты для критичных задач

Тестирование производительности

- Невозможно протестировать всю платформу

- Тесты в dev-окружении

 - › Синтетика – dfsio, terasort

 - › Тесты для критичных задач

- Baseline в тестовом окружении

 - › Нагрузка на кластер

 - › KPI сервисов-клиентов

Тестирование производительности

Невозможно протестировать всю платформу

Тесты в dev-окружении

- › Синтетика – dfsio, terasort

- › Тесты для критичных задач

Baseline в тестовом окружении

- › Нагрузка на кластер

- › KPI сервисов-клиентов

Baseline в боевом окружении

Проблема 3: управление ресурсами

Проблема 3: управление ресурсами

Ограничение ресурсов:

- › Память
- › vCpu

Проблема 3: управление ресурсами

Ограничение ресурсов:

- › Память
- › vCpu



Проблема 3: управление ресурсами

Ограничение ресурсов:

- › Память
- › vCpu



Проблема 3: управление ресурсами

Ограничение ресурсов:

- › Память
- › vCpu



Проблема 3: управление ресурсами

Ограничение ресурсов:

- › Память
- › vCpu



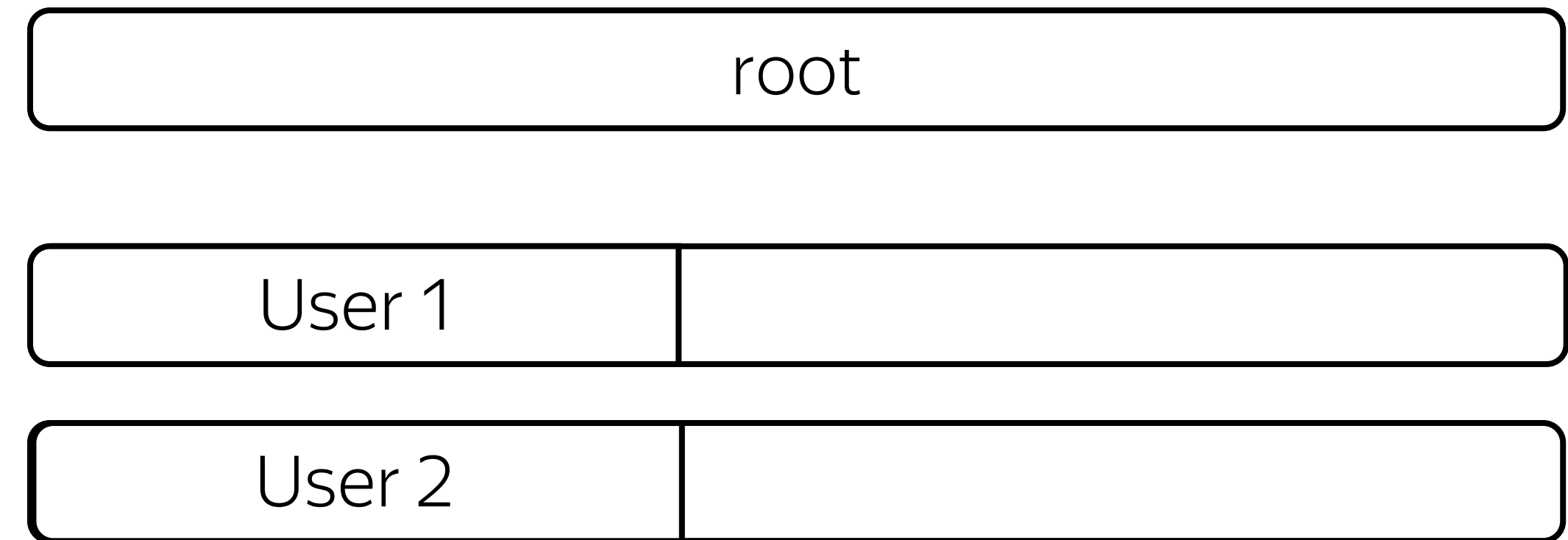
User 1

Проблема 3: управление ресурсами

Ограничение ресурсов:

- › Память
- › vCpu

Иерархические очереди

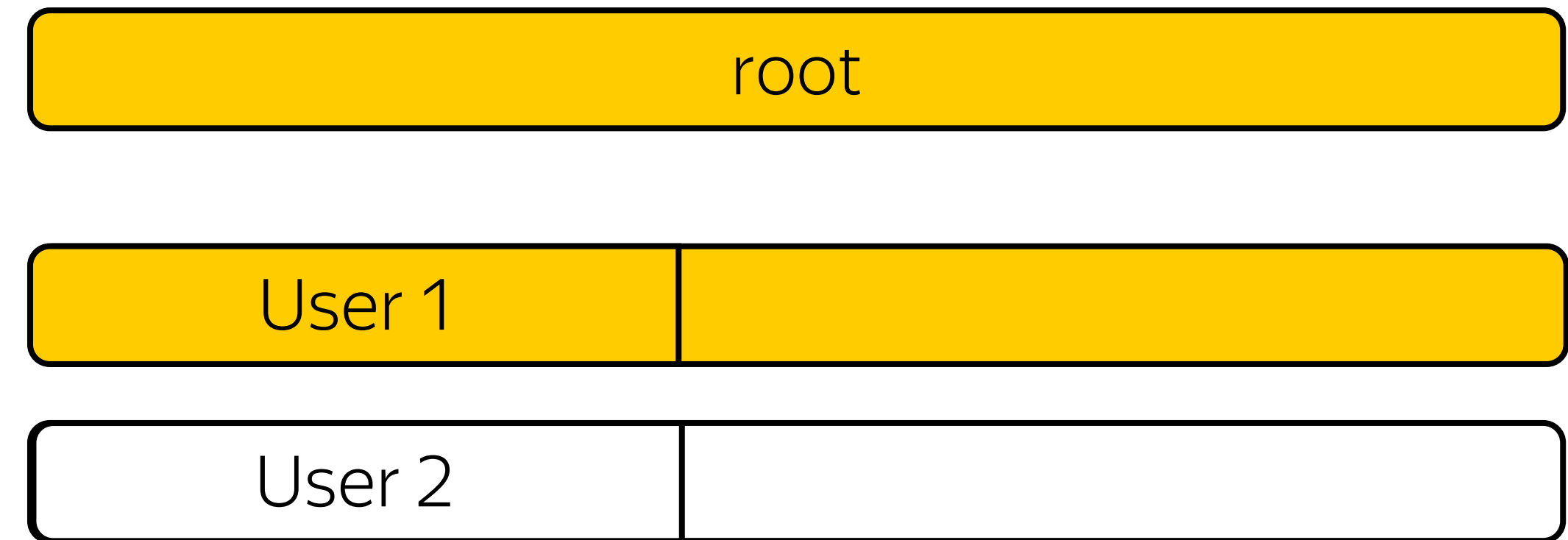


Проблема 3: управление ресурсами

Ограничение ресурсов:

- › Память
- › vCpu

Иерархические очереди



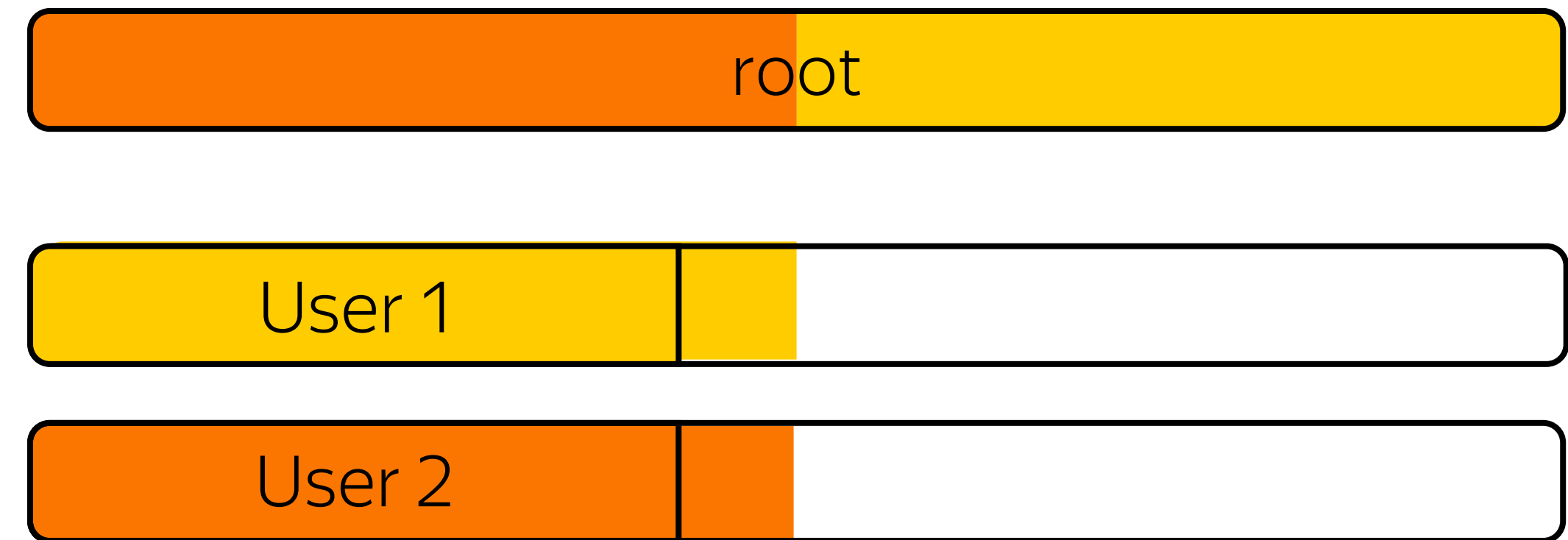
Проблема 3: управление ресурсами

Ограничение ресурсов:

- › Память
- › vCpu

Иерархические очереди

Preemption



Проблема 3: управление ресурсами

Ограничение ресурсов:

- › Память
- › vCpu

Иерархические очереди

Preemption

Fair-scheduler

Strong-soft 80%-20%

root

Soft

Strong

КВОТЫ В HDFS



1 файл – 200 байт в Heart
NameNode

КВОТЫ В HDFS



1 файл – 200 байт в Heart
NameNode



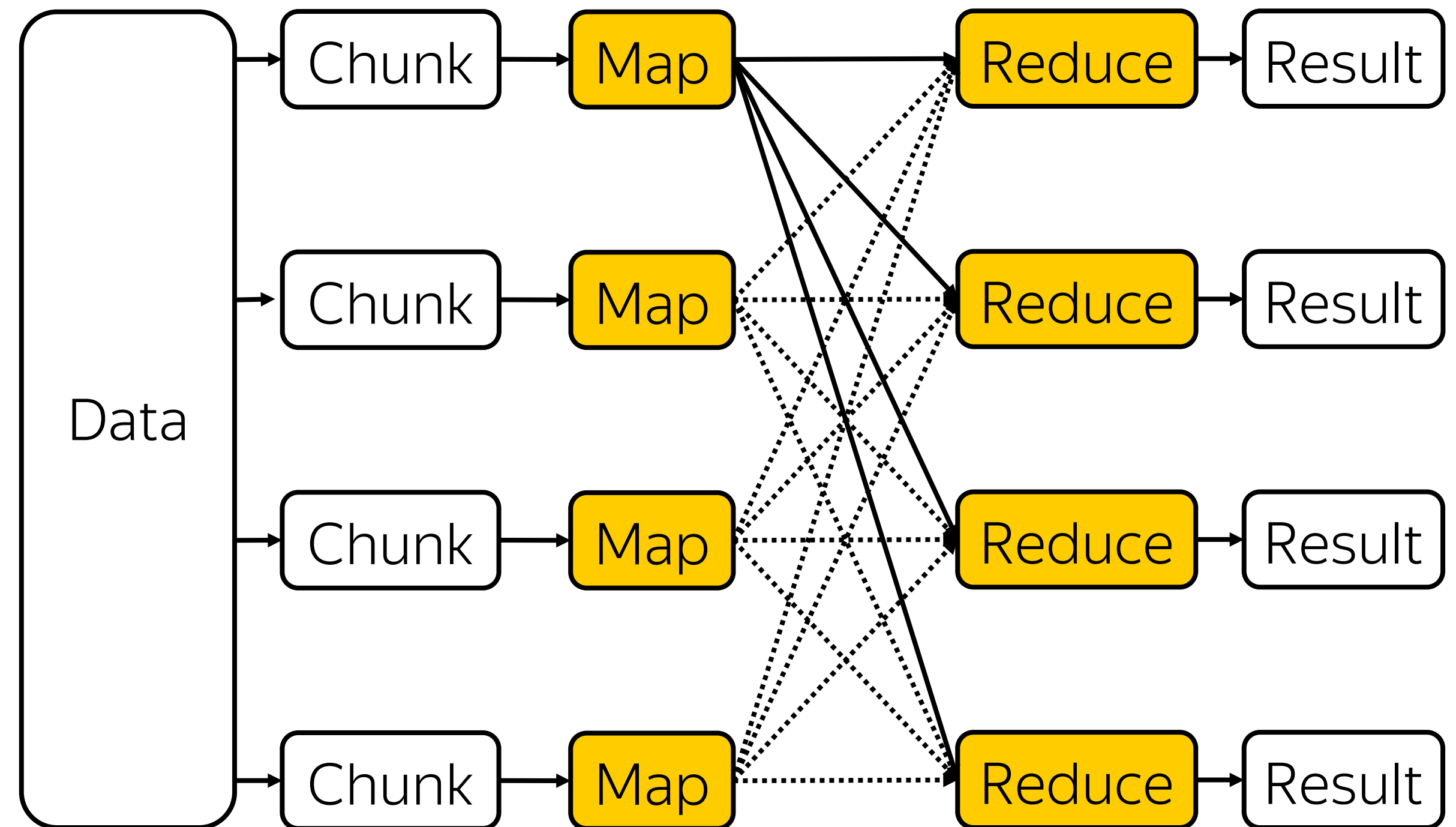
Размер блока – 256 мегабайт

КВОТЫ В HDFS

1 файл – 200 байт в Heap
NameNode

Размер блока – 256 мегабайт

MapReduce создает много
мелких файлов



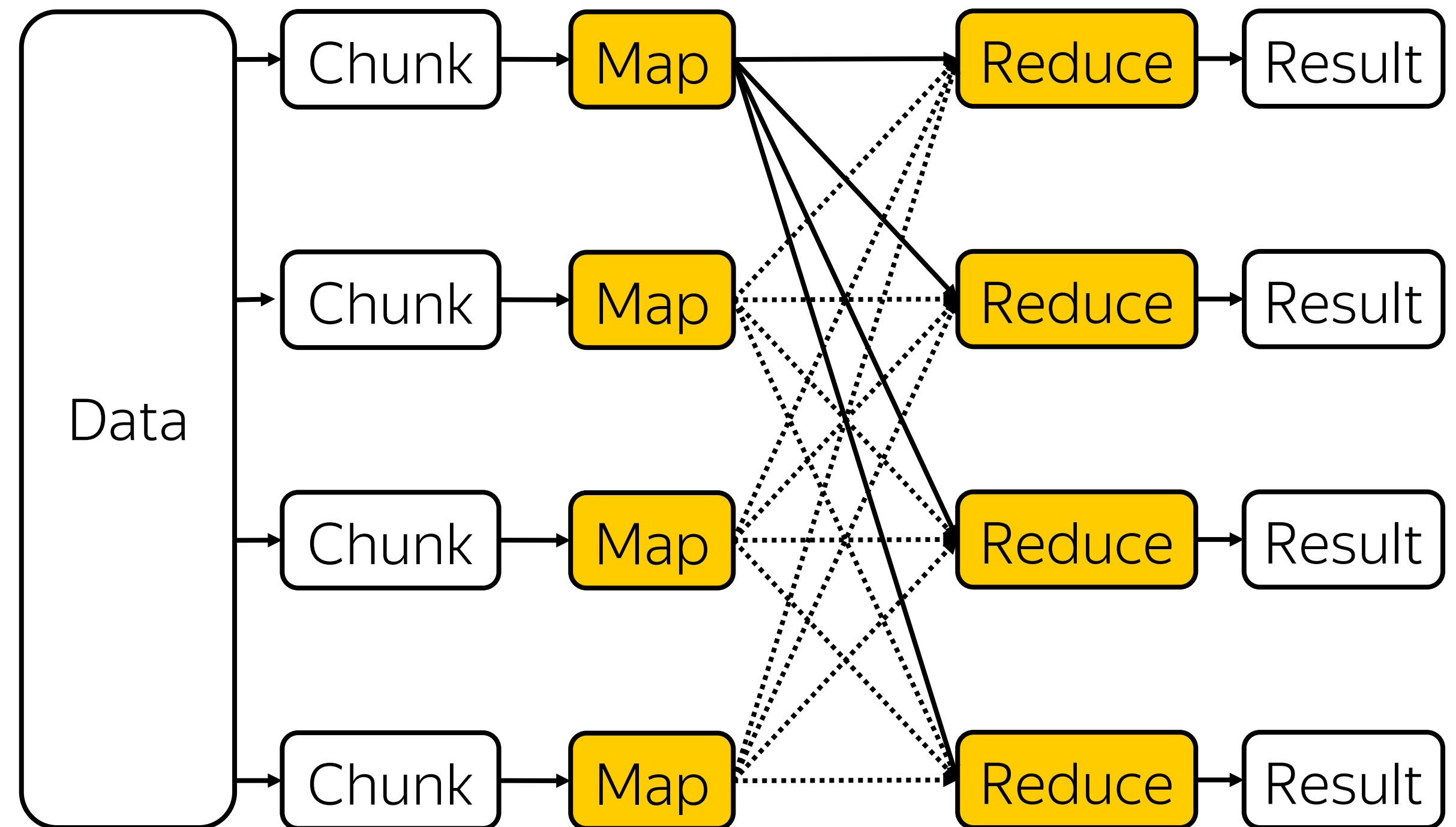
КВОТЫ В HDFS

1 файл – 200 байт в Heap
NameNode

Размер блока – 256 мегабайт

MapReduce создает много
мелких файлов

Квоты на размер и число
файлов



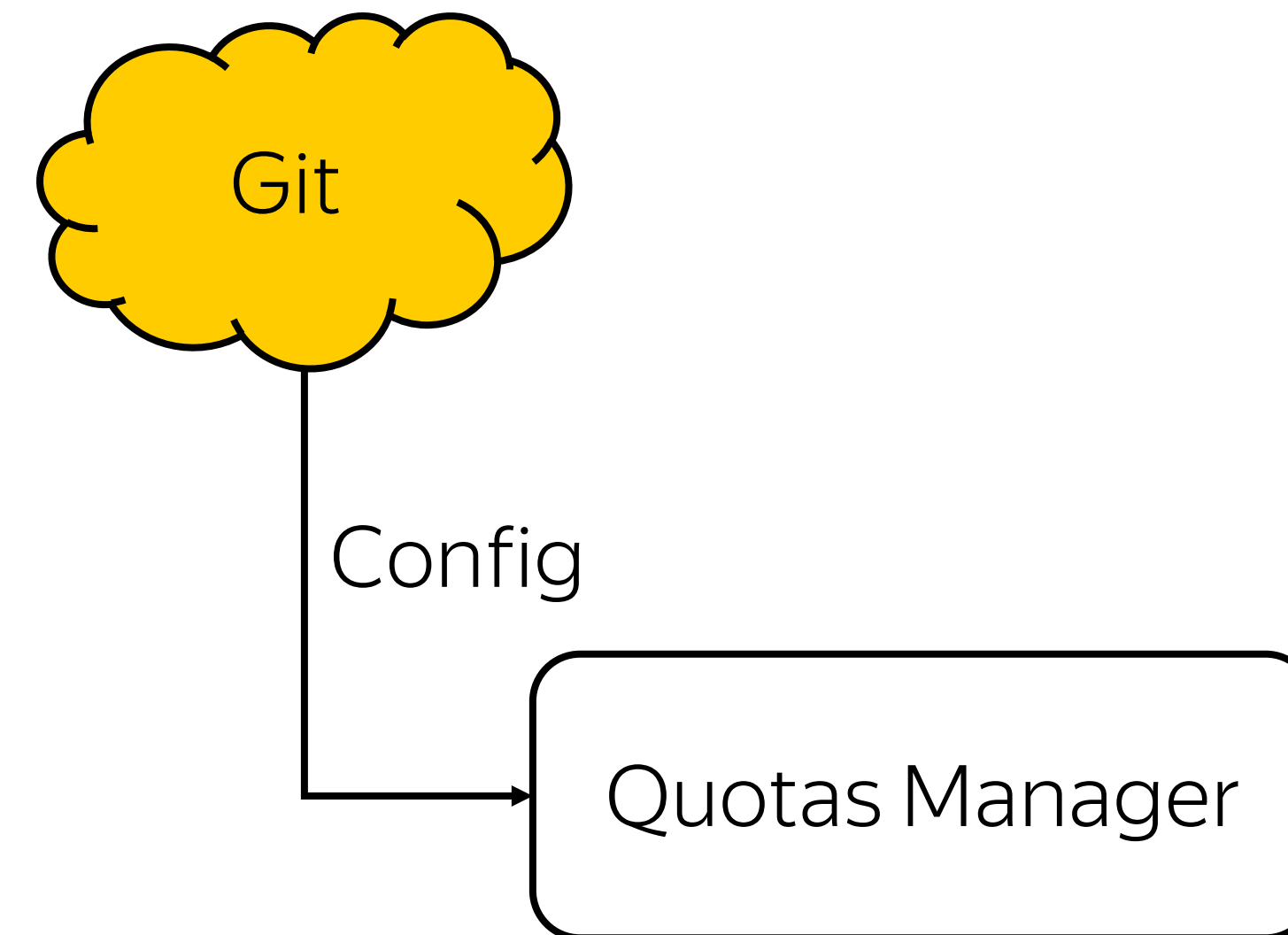
Система управления квотами

Quotas Manager

Система управления квотами

■ Читаемый формат – yaml

› Генерация xml

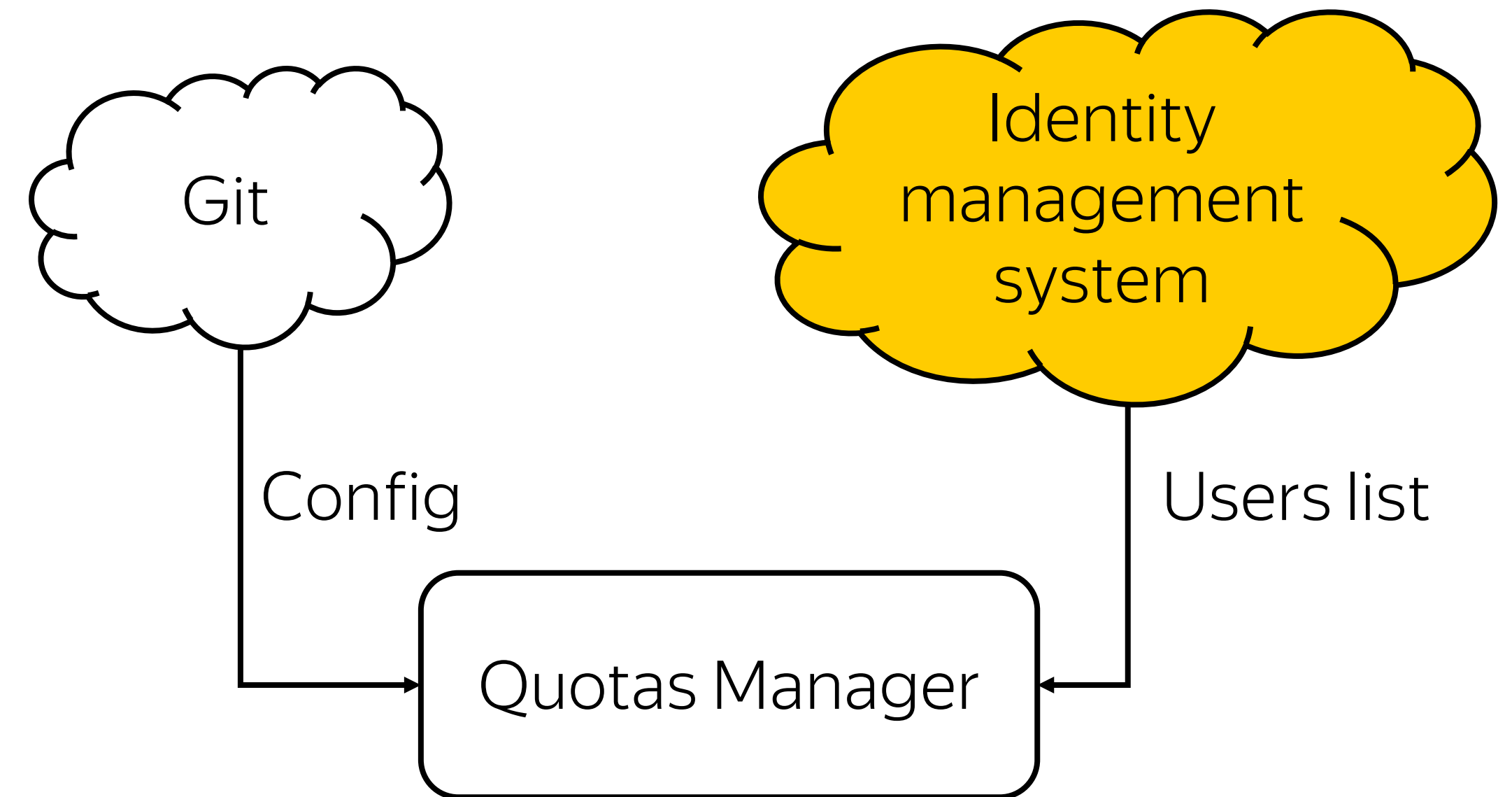


Система управления квотами

■ Читаемый формат – yaml

› Генерация xml

■ Выгрузка данных из IDM



Система управления квотами

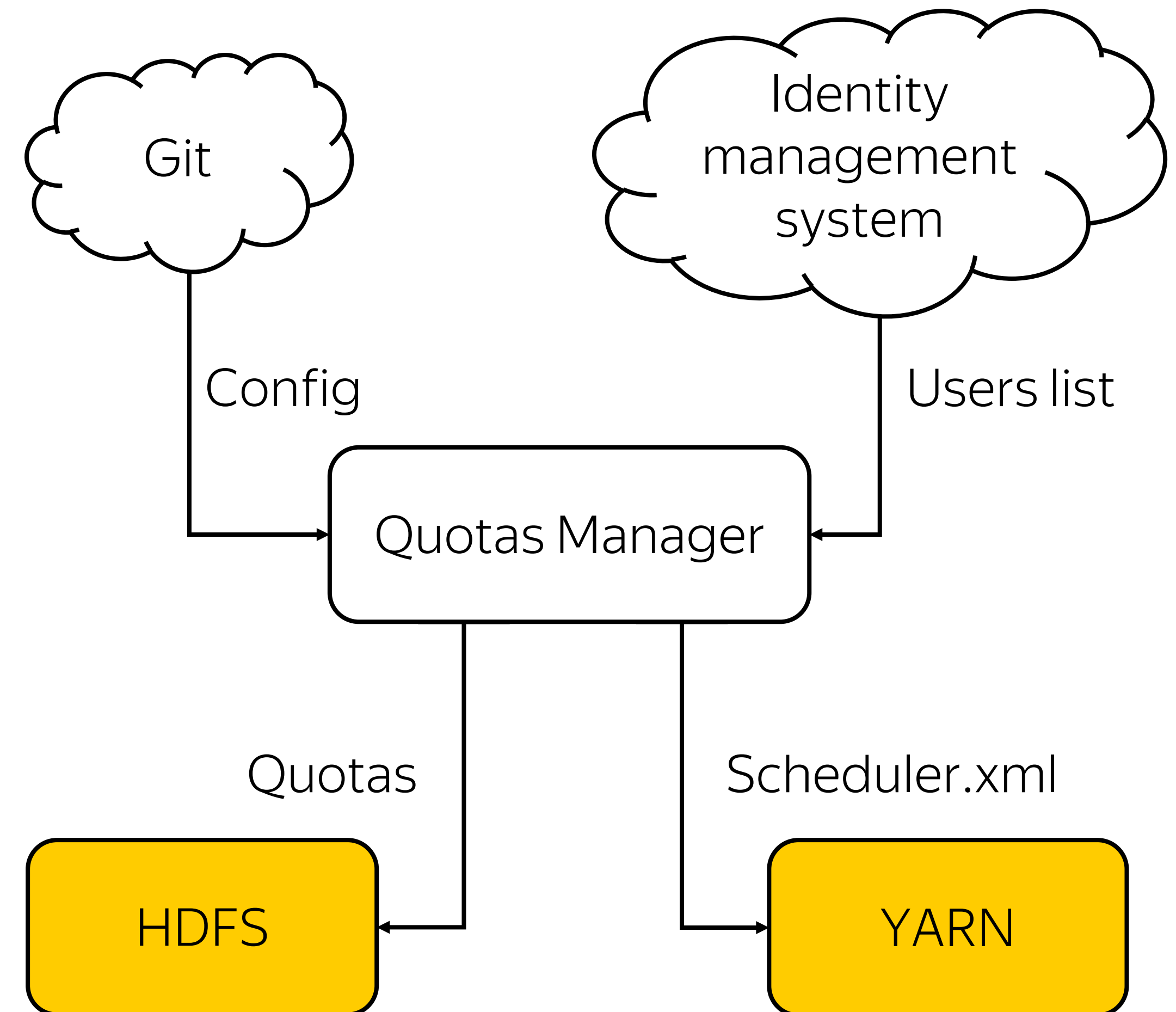
■ Читаемый формат – yaml

› Генерация xml

■ Выгрузка данных из IDM

■ Автоматизация рутины

- › Создание домашних каталогов
- › Квоты на число и объем файлов
- › Очереди в YARN



Проблема 4: копирование данных

Проблема 4: копирование данных

Копирование между кластерами

- › HDFS
- › HBase
- › Hive

Проблема 4: копирование данных

Копирование между кластерами

- › HDFS
- › HBase
- › Hive

Копирование между Hadoop и другими системами

Проблема 4: копирование данных

- Копирование между кластерами

- › HDFS
- › HBase
- › Hive

- Копирование между Hadoop и другими системами

- Рутинная операция

Проблема 4: копирование данных

- Копирование между кластерами

- › HDFS
- › HBase
- › Hive

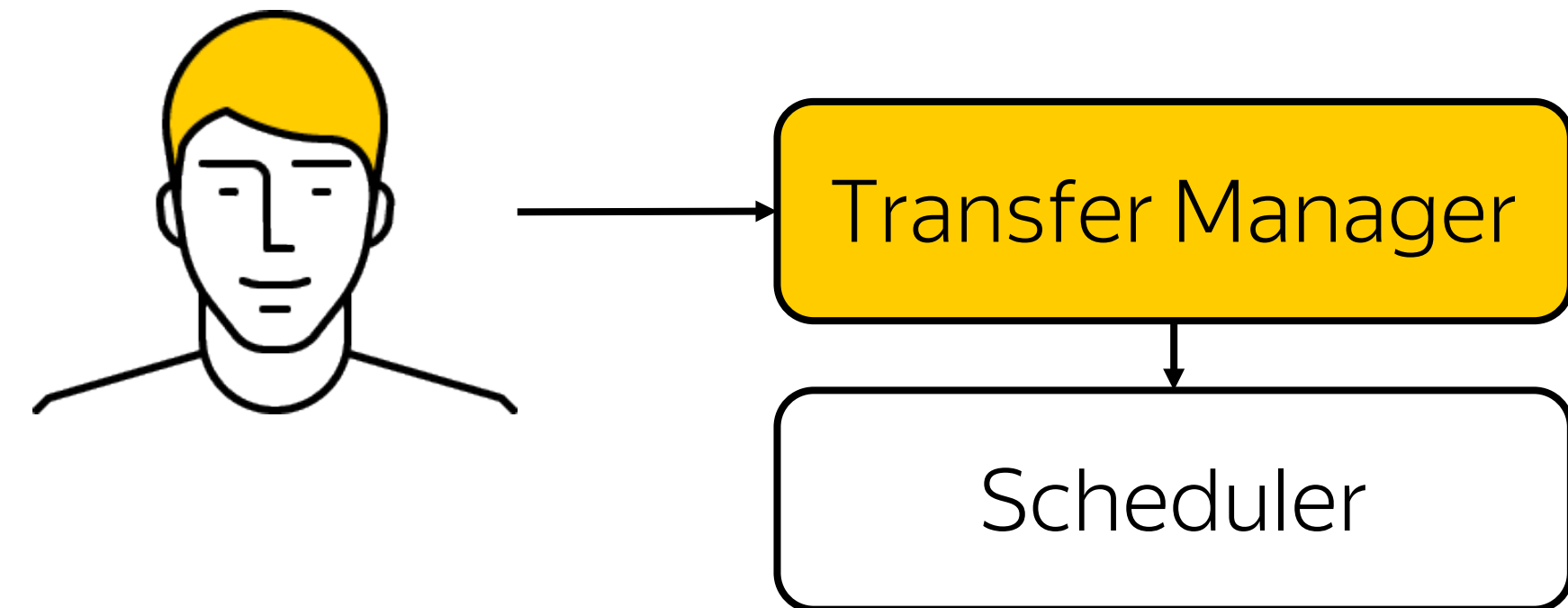
- Копирование между Hadoop и другими системами

- Рутинная операция

- Возможность потерять данные

Transfer Manager

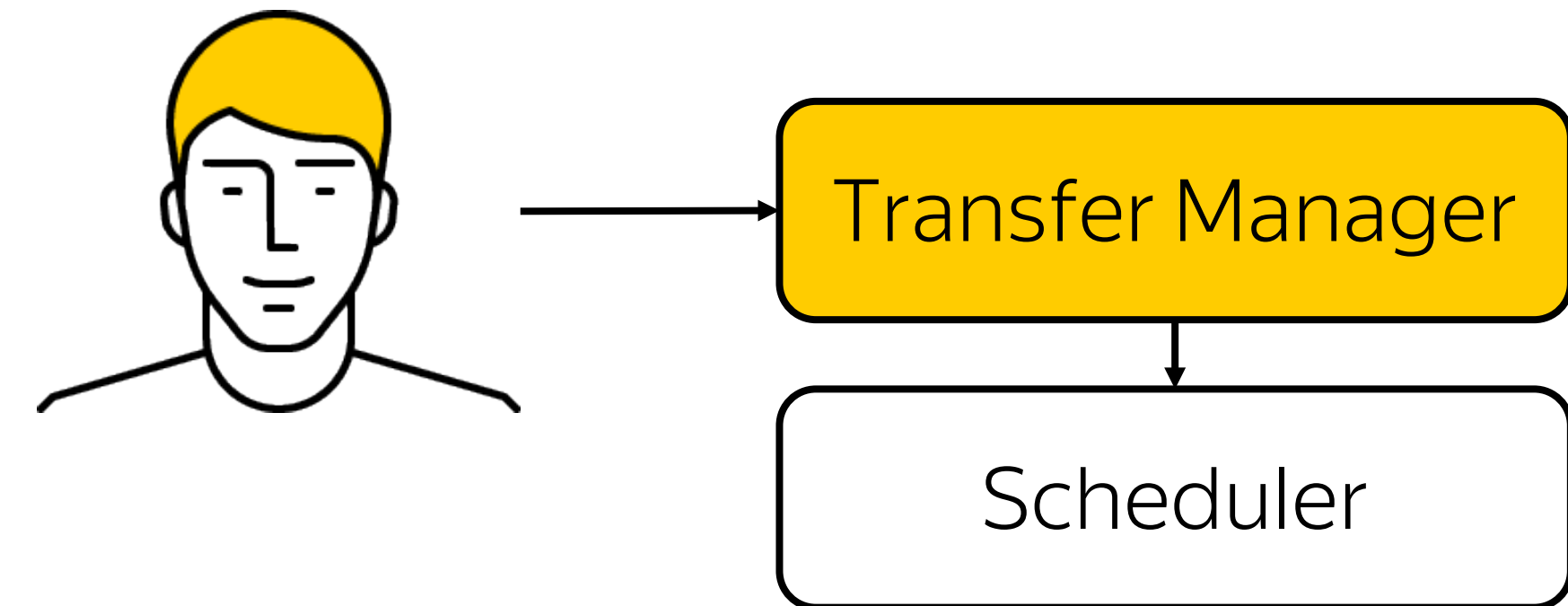
Единый интерфейс пользователя



Transfer Manager

Единый интерфейс пользователя

Правила по направлению
копирования



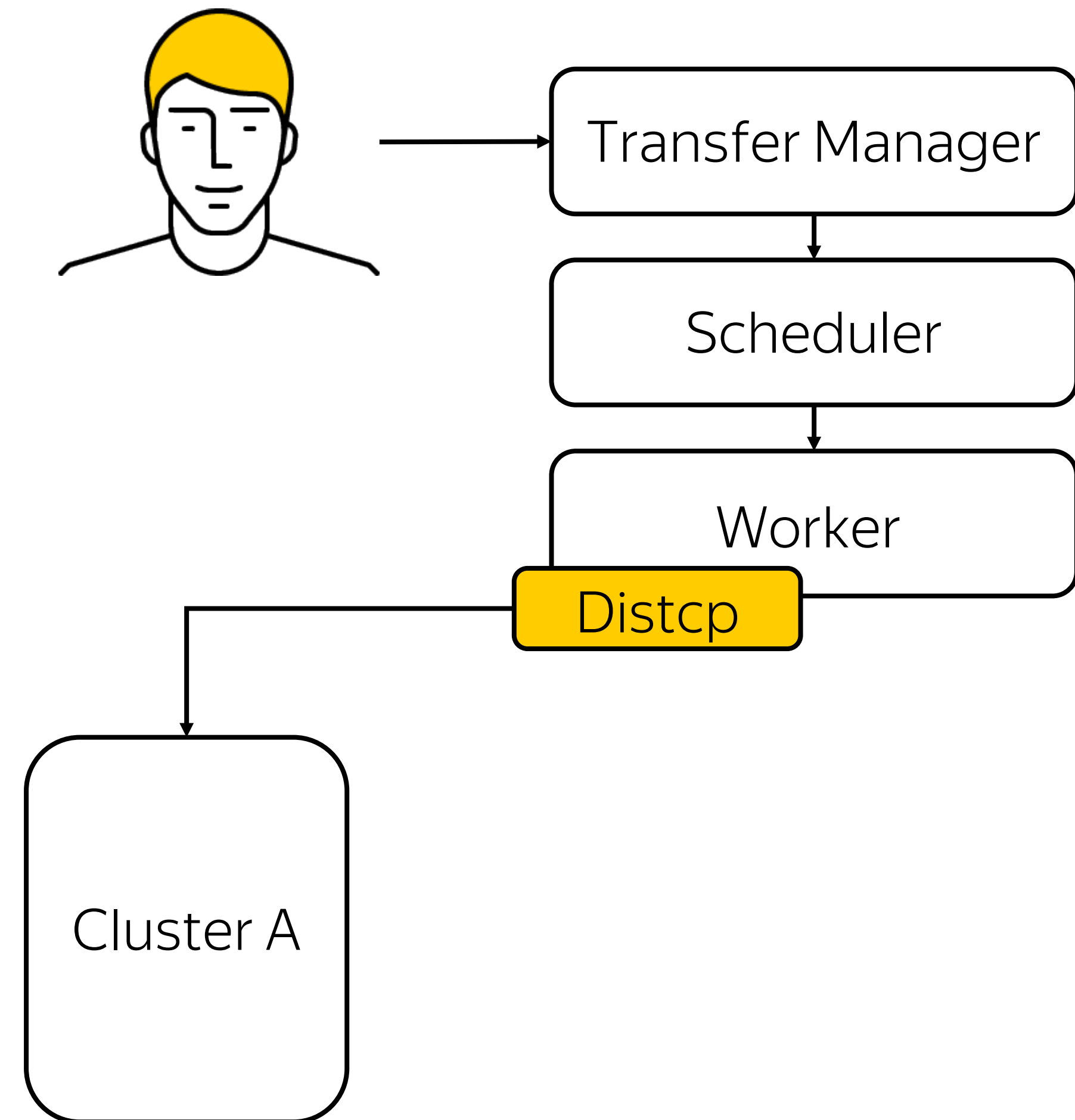
Transfer Manager

Единый интерфейс пользователя

Правила по направлению копирования

Защита от перезаписи

Проверка прав доступа



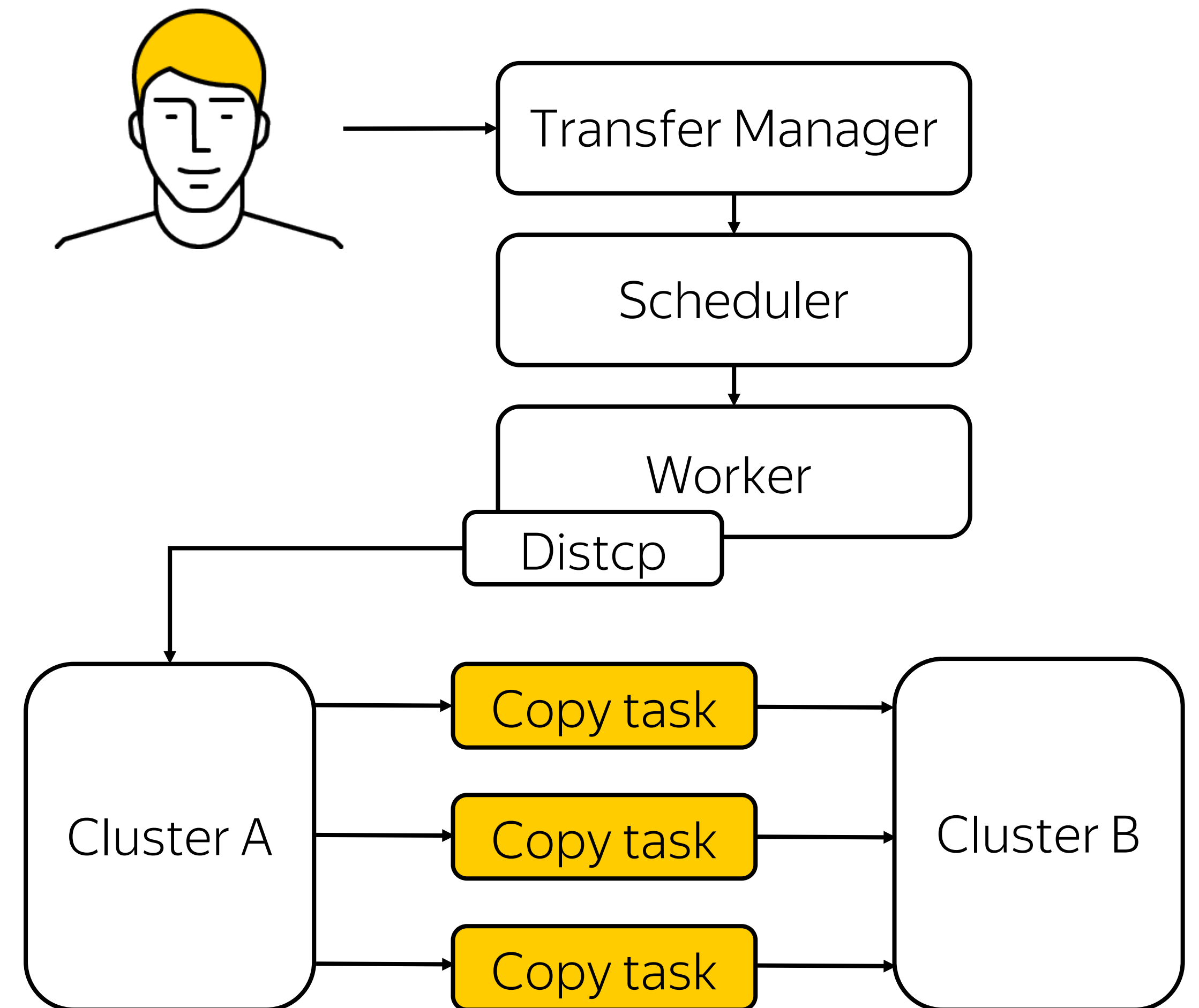
Transfer Manager

Единый интерфейс пользователя

Правила по направлению копирования

Защита от перезаписи

Проверка прав доступа



Проблема 5: отказ hardware

Проблема 5: отказ hardware

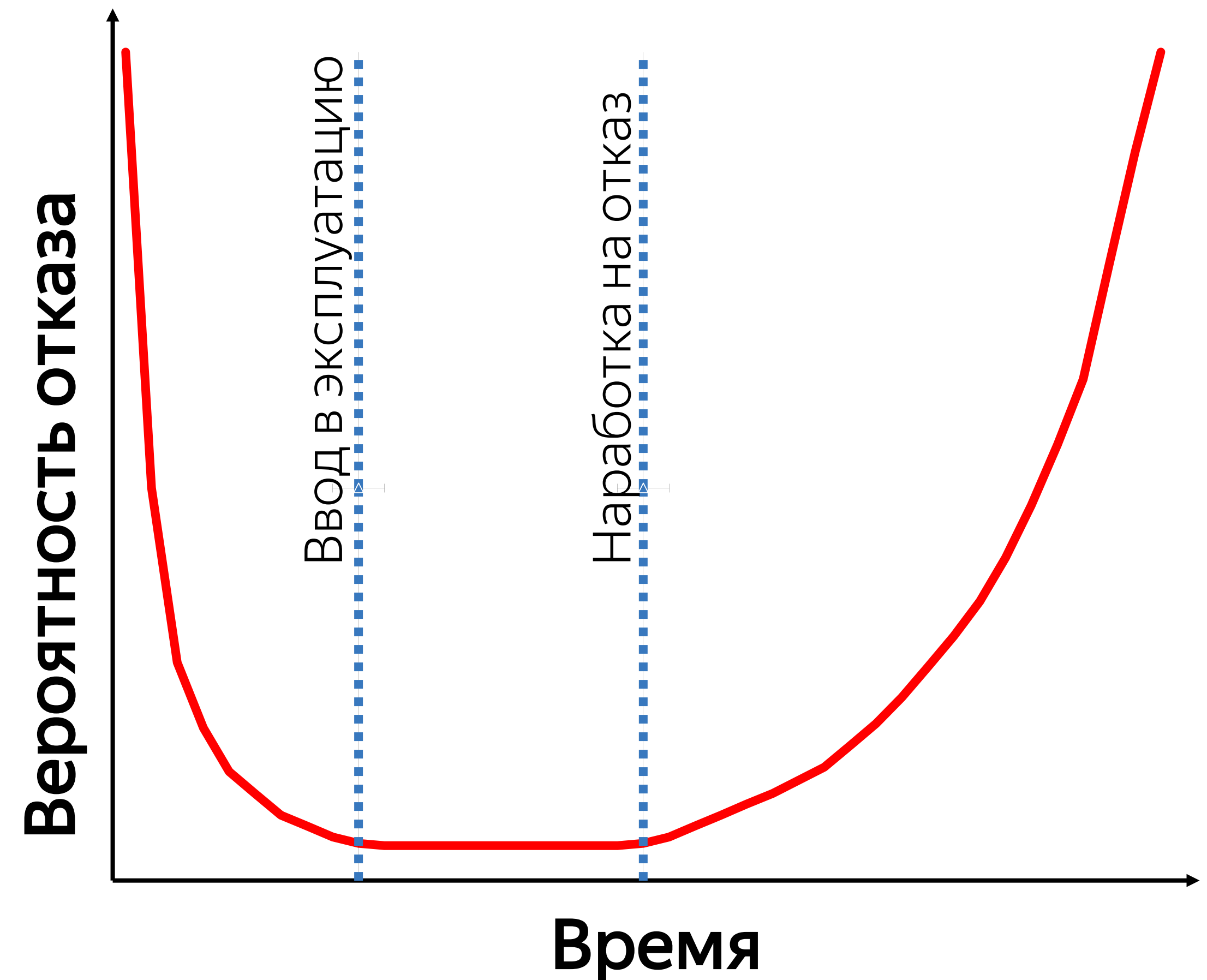
 Сервера выходят из строя

Проблема 5: отказ hardware

- Сервера выходят из строя
- Серверов много

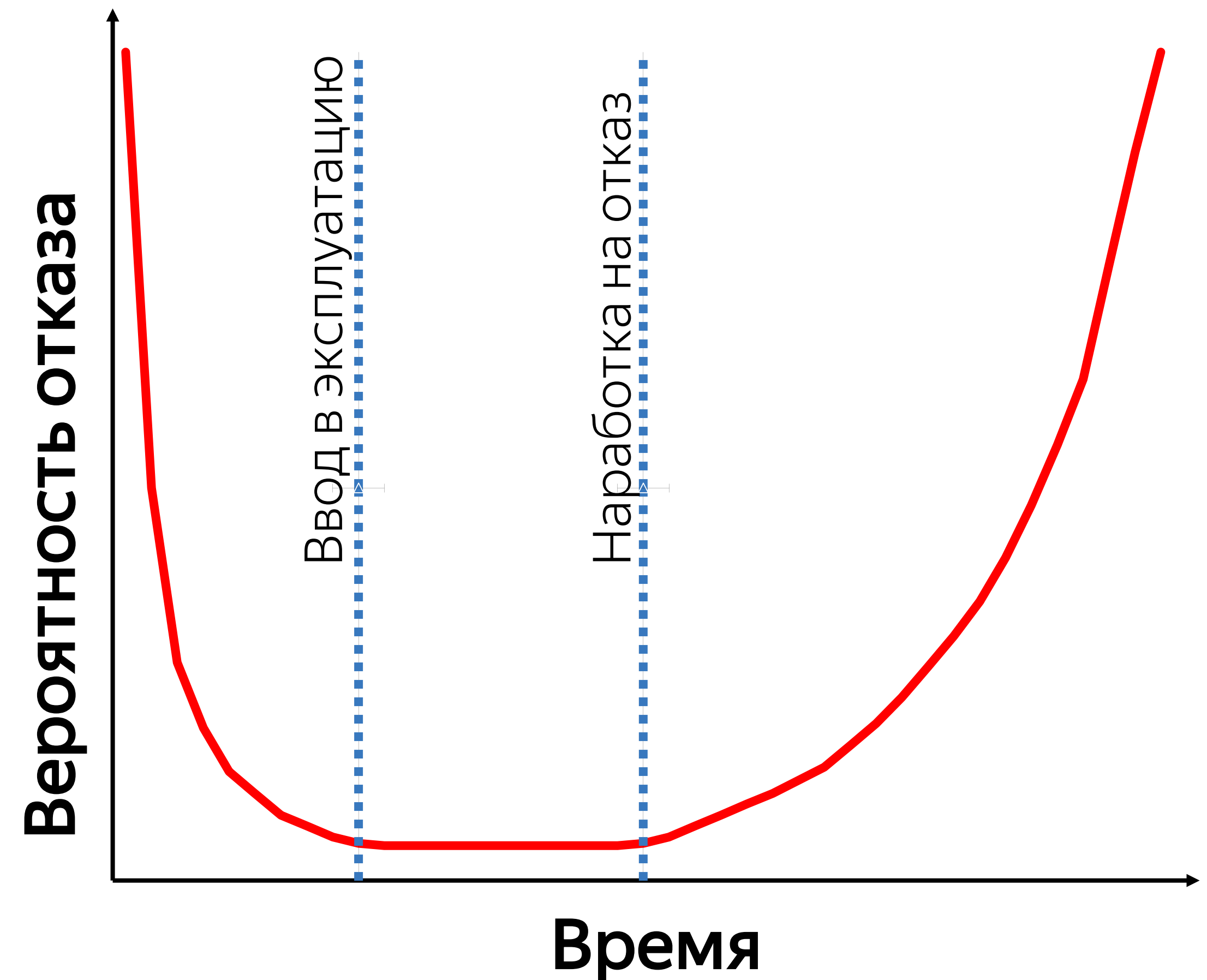
Проблема 5: отказ hardware

- Сервера выходят из строя
- Серверов много
- Рост числа отказов относительно времени работы кластера



Проблема 5: отказ hardware

- Сервера выходят из строя
- Серверов много
- Рост числа отказов относительно времени работы кластера
- Можно потерять данные



Проблема 5: отказ hardware

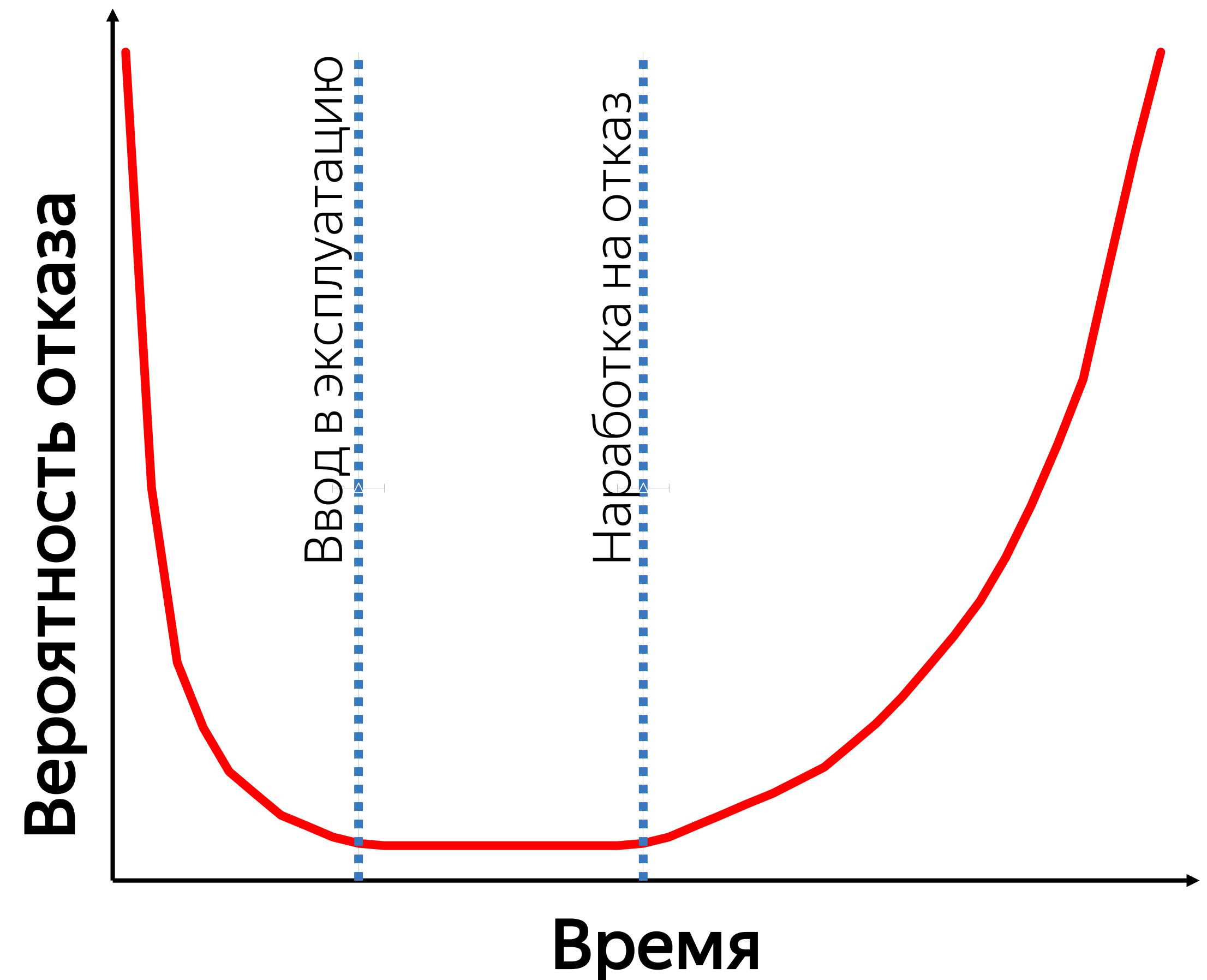
Сервера выходят из строя

Серверов много

Рост числа отказов относительно времени работы кластера

Можно потерять данные

Рутина - съедает время



Автоматическая замена дисков

 hw_watch

› Мониторинг HDD, RAM, Network

Автоматическая замена дисков

hw_watcher

- › Мониторинг HDD, RAM, Network
- › Скрипты для декомиссии сервера из кластера

Автоматическая замена дисков

hw_watcher

- › Мониторинг HDD, RAM, Network
- › Скрипты для декомиссии сервера из кластера
- › Создание заявки на замену

Автоматическая замена дисков

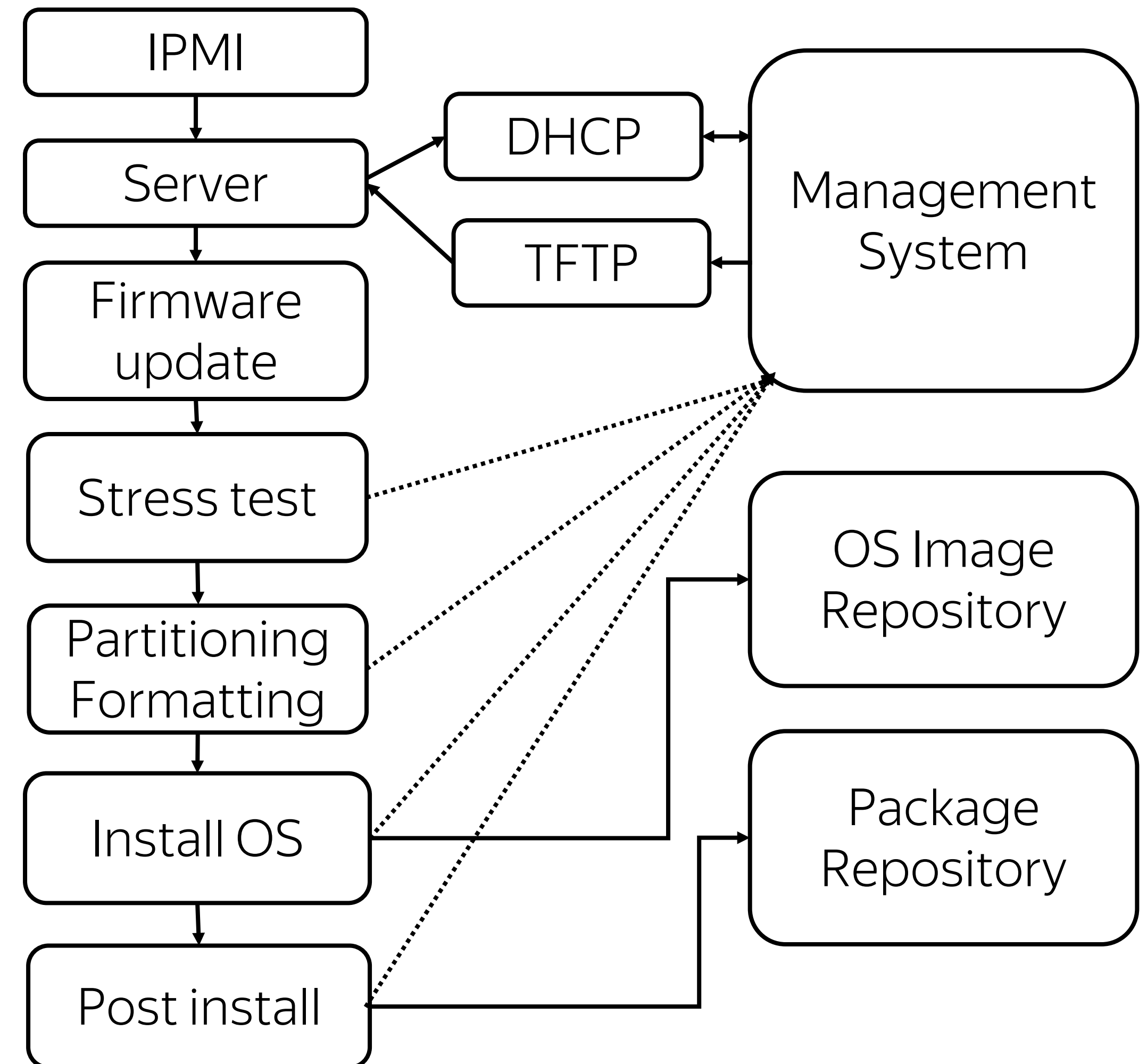
hw_watcher

- › Мониторинг HDD, RAM, Network
- › Скрипты для декомиссии сервера из кластера
- › Создание заявки на замену
- › Инициализация разделов и файловых систем после замены

Bootstrap

Автоматические:

- › Стресс-тестирование
- › Разметка и форматирование дисков
- › Установка OS и базовых пакетов
- › Установка и конфигурация salt-minion

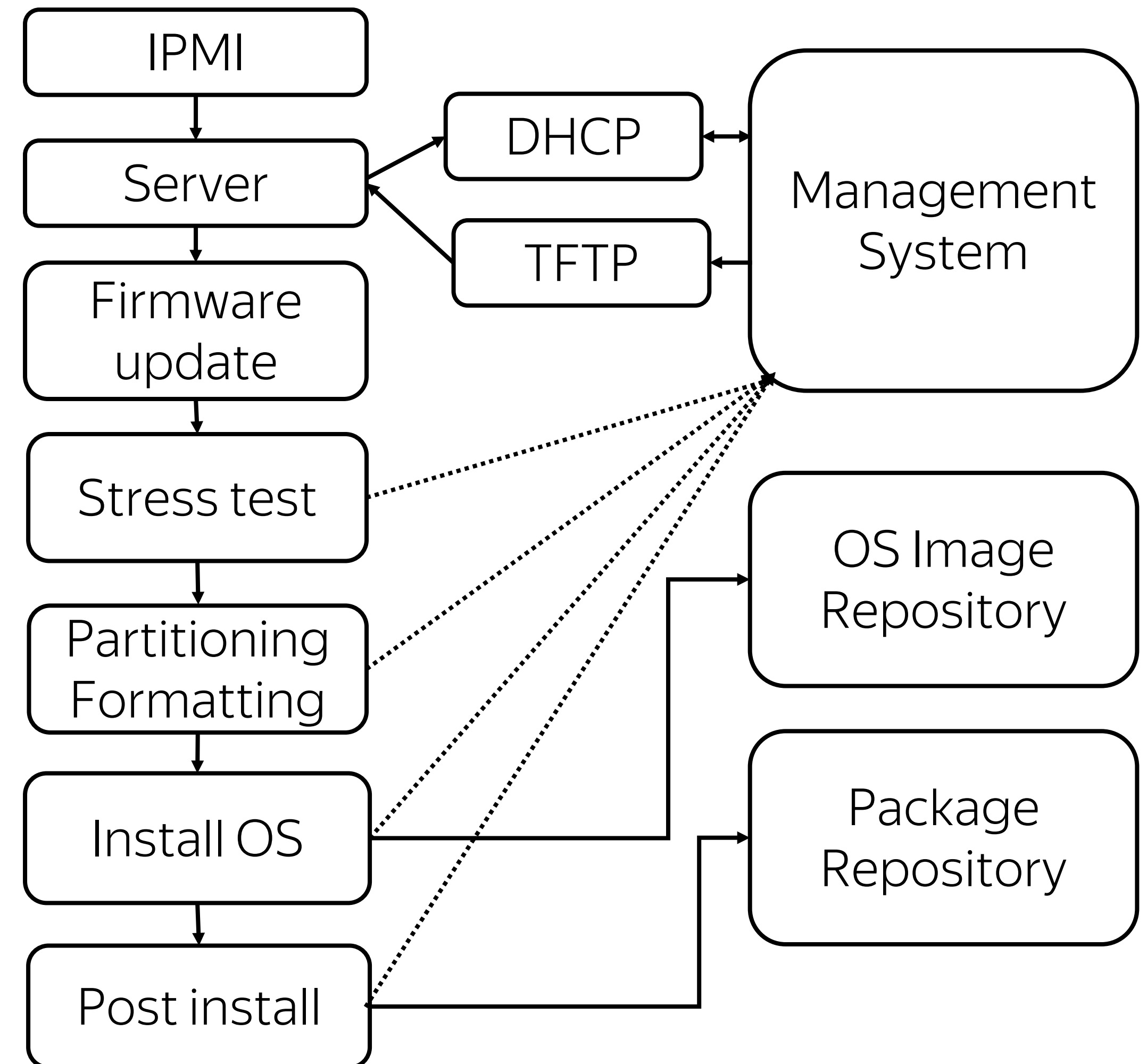


Bootstrap

Автоматические:

- › Стресс-тестирование
- › Разметка и форматирование дисков
- › Установка OS и базовых пакетов
- › Установка и конфигурация salt-minion

Переналивка сервера без
потери информации



Проблема 6: Что происходит?

Проблема 6: Что происходит?

■ Ответы на вопросы пользователей:

› Кто занял кластер?

Проблема 6: Что происходит?

■ Ответы на вопросы пользователей:

- › Кто занял кластер?
- › Почему задачи работают медленно?

Проблема 6: Что происходит?

■ Ответы на вопросы пользователей:

- › Кто занял кластер?
- › Почему задачи работают медленно?
- › Сколько ресурсов потребляет наш сервис?

Проблема 6: Что происходит?

■ Ответы на вопросы пользователей:

- › Кто занял кластер?
- › Почему задачи работают медленно?
- › Сколько ресурсов потребляет наш сервис?
- › Какие параметры запуска задач менять?

Проблема 6: Что происходит?

■ Ответы на вопросы пользователей:

- › Кто занял кластер?
- › Почему задачи работают медленно?
- › Сколько ресурсов потребляет наш сервис?
- › Какие параметры запуска задач менять?
- › Как изменения повлияли на задачи?

Dr.Elephant – анализ и проверка задач

Search

Job/App ID

Job ID

Flow Execution URL/ID

Flow Exec URL/ID

User

User

Queue

Queue

☐ Job Type

Spark

☒ Severity

Critical

All Analysis

☐ Job Finish Date

From: mm/dd/yyyy

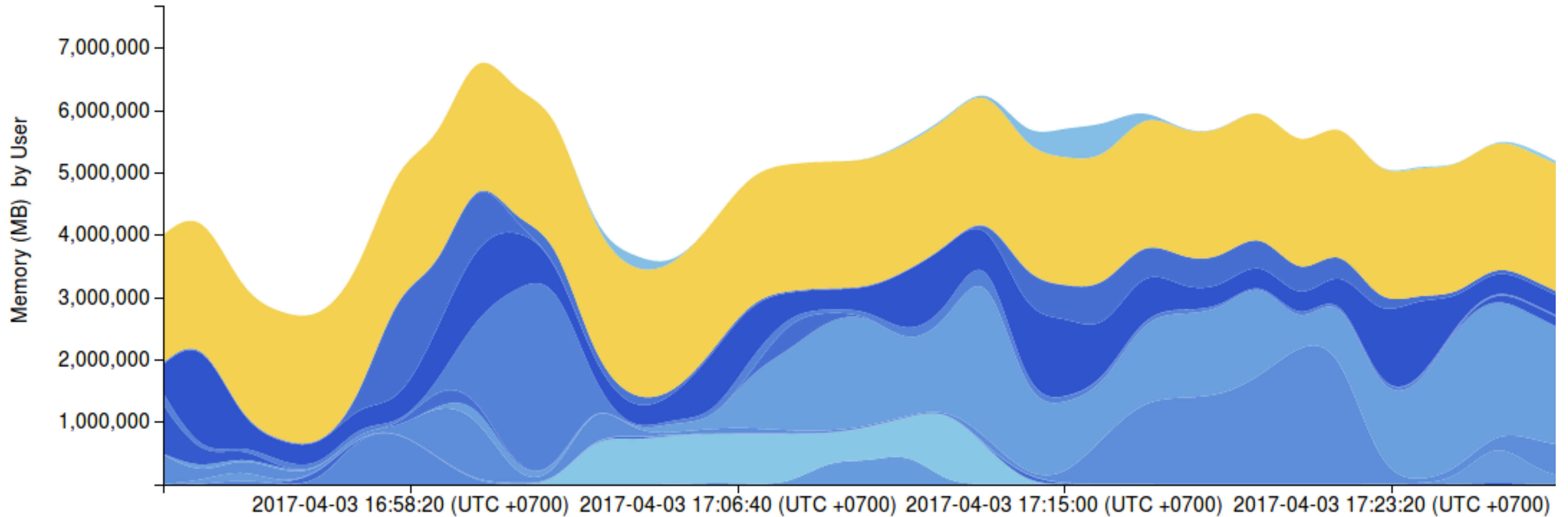
To: mm/dd/yyyy

Get Result

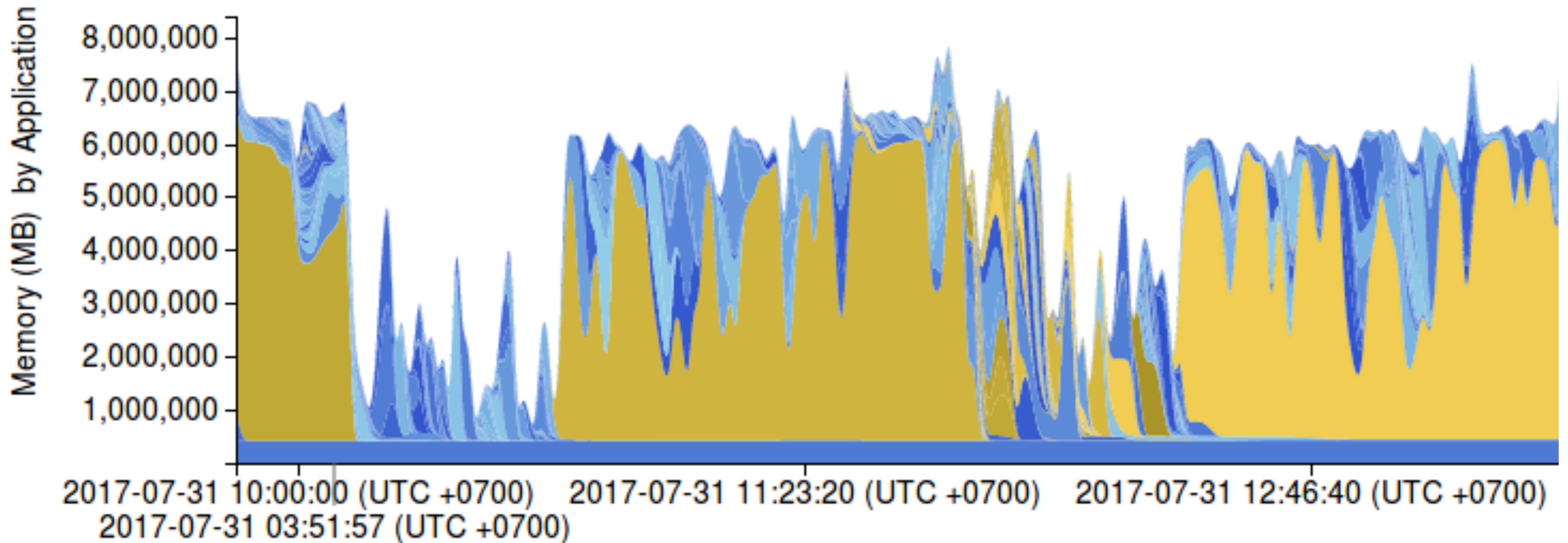
Results: Showing 20 of 2,225,478			
[robot-market-calc-ts] [Hive] application_1510569628086_565024 Mon Feb 05 2018 16:43:07 GMT+0700 (+07)			
INSERT INTO TABLE marketpage...gen=1085619))(Stage-1)			
Mapper Data Skew	Mapper GC	Mapper Time	Mapper Speed
Mapper Spill	Mapper Memory	Reducer Data Skew	Reducer GC
Reducer Time	Reducer Memory	Shuffle & Sort	Distributed Cache Limit
0.113 GB Hours	0 %	0:00:17	13.04 %
[robot-market-exp-ts] [Hive] application_1510569628086_565023 Mon Feb 05 2018 16:42:49 GMT+0700 (+07)			
insert into table geo_cat...a.categ_id(Stage-4)			
Mapper Data Skew	Mapper GC	Mapper Time	Mapper Speed
Mapper Spill	Mapper Memory	Reducer Data Skew	Reducer GC
Reducer Time	Reducer Memory	Shuffle & Sort	Distributed Cache Limit
0.004 GB Hours	2.98 %	0:00:15	14.32 %
[robot-market-calc-ts] [Hive] application_1510569628086_565022 Mon Feb 05 2018 16:42:46 GMT+0700 (+07)			
INSERT INTO TABLE marketpage...gen=1085619))(Stage-1)			
Mapper Data Skew	Mapper GC	Mapper Time	Mapper Speed
Mapper Spill	Mapper Memory	Reducer Data Skew	Reducer GC
Reducer Time	Reducer Memory	Shuffle & Sort	Distributed Cache Limit
0.077 GB Hours	0 %	0:00:12	17.13 %
[robot-market-calc-ts] [Hive] application_1510569628086_565021 Mon Feb 05 2018 16:42:31 GMT+0700 (+07)			
INSERT INTO TABLE marketpage...gen=1085619))(Stage-1)			
Mapper Data Skew	Mapper GC	Mapper Time	Mapper Speed
Mapper Spill	Mapper Memory	Reducer Data Skew	Reducer GC
Reducer Time	Reducer Memory	Shuffle & Sort	Distributed Cache Limit
0.147 GB Hours	0 %	0:00:15	14.88 %
[robot-market-exp-ts] [Hive] application_1510569628086_565020 Mon Feb 05 2018 16:42:26 GMT+0700 (+07)			
insert into table geo_cat...a.categ_id(Stage-2)			
Mapper Data Skew	Mapper GC	Mapper Time	Mapper Speed
Mapper Spill	Mapper Memory	Reducer Data Skew	Reducer GC
Reducer Time	Reducer Memory	Shuffle & Sort	Distributed Cache Limit
0.095 GB Hours	0 %	0:00:18	23.32 %

[robot-mrk-vis-clust] [Unknown] application_1510569628086_564938 Mon Feb 05 2018 16:18:56 GMT+0700 (+07)	
generation-divider-job	
Jobtracker: http://m02.cavy.hdp.yandex.net:8088/proxy/application_1510569628086_564938/jobhistory/job/job_1510569628086_564938	
Mapper Data Skew	Mapper GC
Mapper Time	Mapper Speed
Mapper Spill	Mapper Memory
Reducer Data Skew	Reducer GC
Reducer Time	Reducer Memory
Shuffle & Sort	Distributed Cache Limit
24.794 GB Hours	0 %
0:04:44	3.30 %
[Explain]	
Mapper Data Skew	
Severity: None	
Group A	1 tasks @ 163 MB avg
Group B	1172 tasks @ 256 MB avg
Number of tasks	1173
Mapper GC	
Severity: None	
Avg task CPU time (ms)	9136
Avg task GC time (ms)	722
Avg task runtime (ms)	30378
Number of tasks	1173
Task GC/CPU ratio	0.07902802101576183
Mapper Time	
Severity: Critical [Explain]	
Average task input size	255 MB
Average task runtime	30 sec
Max task runtime	1 min 6 sec
Min task runtime	9 sec
Number of tasks	1173
Mapper Speed	

Inviso – потребляемые задачами ресурсы



Inviso – потребляемые задачами ресурсы



Проблема 7: рост нагрузки и деградация

Проблема 7: рост нагрузки и деградация

Сервисы развиваются

- › Растет объем данных
- › Растет время работы задач
- › Появляются новые компоненты

Проблема 7: рост нагрузки и деградация

Сервисы развиваются

- › Растет объем данных
- › Растет время работы задач
- › Появляются новые компоненты

При использовании 90% HDFS производительность деградирует

При использовании 95% HDFS кластер перестает работать

Проблема 7: рост нагрузки и деградация

Сервисы развиваются

- › Растет объем данных
- › Растет время работы задач
- › Появляются новые компоненты

При использовании 90% HDFS производительность деградирует

При использовании 95% HDFS кластер перестает работать

Разнородные машины в кластере вызывают проблемы из-за неоднородной загрузки

Retention policy & capacity planning

 Балансировка нагрузки

Retention policy & capacity planning

- Балансировка нагрузки
- Однородное окружение

Retention policy & capacity planning

■ Балансировка нагрузки

■ Однородное окружение

■ Retention policy

- › Квоты на размер, число файлов
- › Слияние мелких файлов
- › Очистка - /var, /tmp

Retention policy & capacity planning

■ Балансировка нагрузки

■ Однородное окружение

■ Retention policy

- › Квоты на размер, число файлов
- › Слияние мелких файлов
- › Очистка - /var, /tmp

■ Capacity planning

- › Известные планы по развитию сервисов
- › Отчеты по потребляемым ресурсам по пользователям
- › Дашборд с прогнозом потребления через X месяцев
- › Запас на крайний случай

Проблема 8: потеря данных

Проблема 8: потеря данных

Hardware:

- › Blackout датацентра
- › Отказ стойки или коммутатора

Проблема 8: потеря данных

Hardware:

- › Blackout датацентра
- › Отказ стойки или коммутатора

Логические:

- › Drop if exists

Проблема 8: потеря данных

Hardware:

- › Blackout датацентра
- › Отказ стойки или коммутатора

Логические:

- › Drop if exists

Человеческий фактор

Как не потерять данные

■ Hadoop - отказоустойчивая система

- › Правильная расстановка машин
- › Включенный Rack-awareness
- › HA для всех критичных сервисов

Как не потерять данные

■ Hadoop - отказоустойчивая система

- › Правильная расстановка машин
- › Включенный Rack-awareness
- › HA для всех критичных сервисов

■ Защита от пользователей:

- › Включенный Trash
- › Ограничение прав
- › Отдельные учетные записи для пользователей и процессов

Заключение

 Примеры сервисов на основе Hadoop

Заключение

- Примеры сервисов на основе Hadoop
- Какие задачи решает собственная сборка Hadoop

Заключение

- Примеры сервисов на основе Hadoop
- Какие задачи решает собственная сборка Hadoop
- Удобная инфраструктура для автоматизации задач пользователей

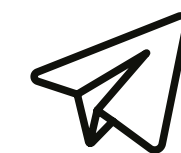
Спасибо за внимание!

Игорь Андреев

Системный администратор



iandreyev@yandex-team.ru



[@iandreyev](https://t.me/iandreyev)



8-923-184-18-75

Bonus



Yandex Tables – сбор и хранение метрик

Yandex Tables – сбор и хранение метрик

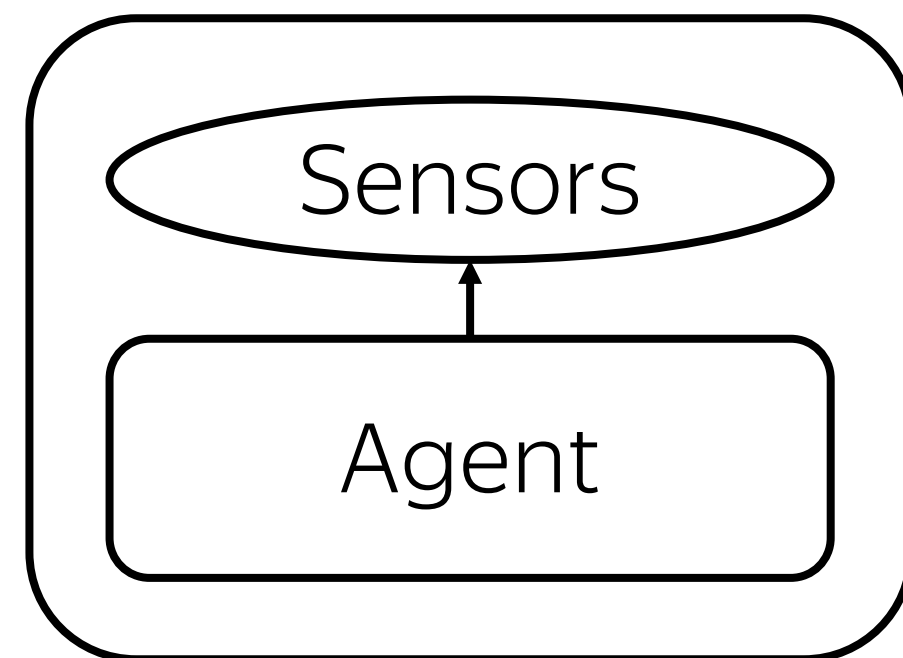
 <https://github.com/yandex/opentsdb-flume>

- › Apache Flume
- › Opentsdb

Yandex Tables – сбор и хранение метрик

<https://github.com/yandex/opentsdb-flume>

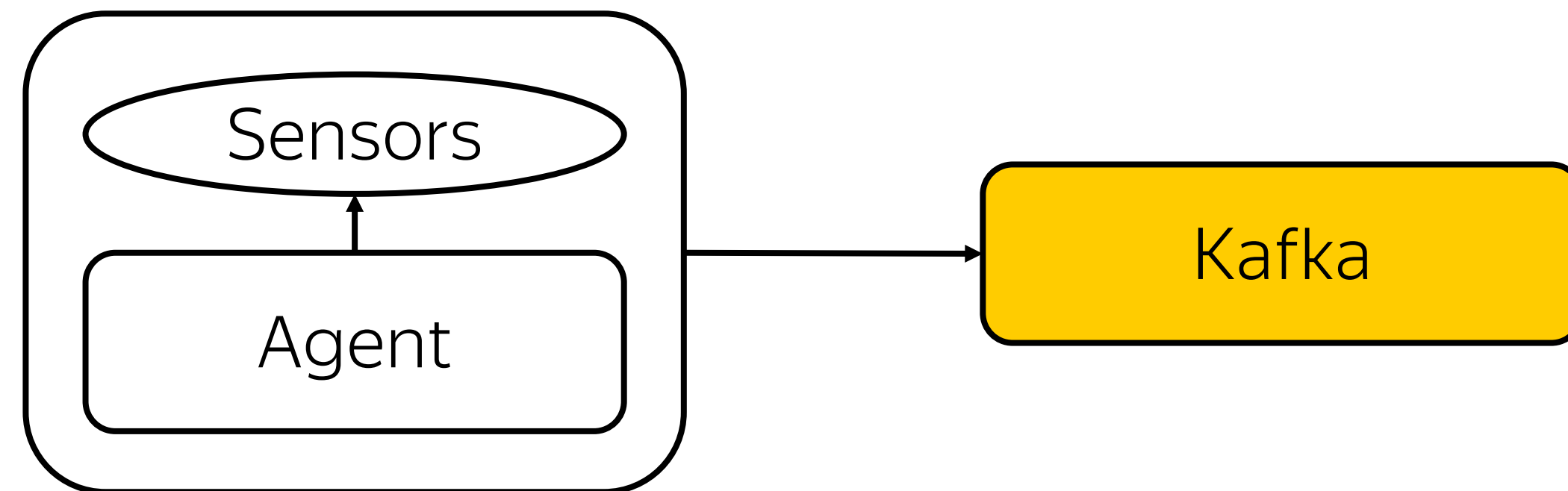
- › Apache Flume
- › Opentsdb



Yandex Tables – сбор и хранение метрик

<https://github.com/yandex/opentsdb-flume>

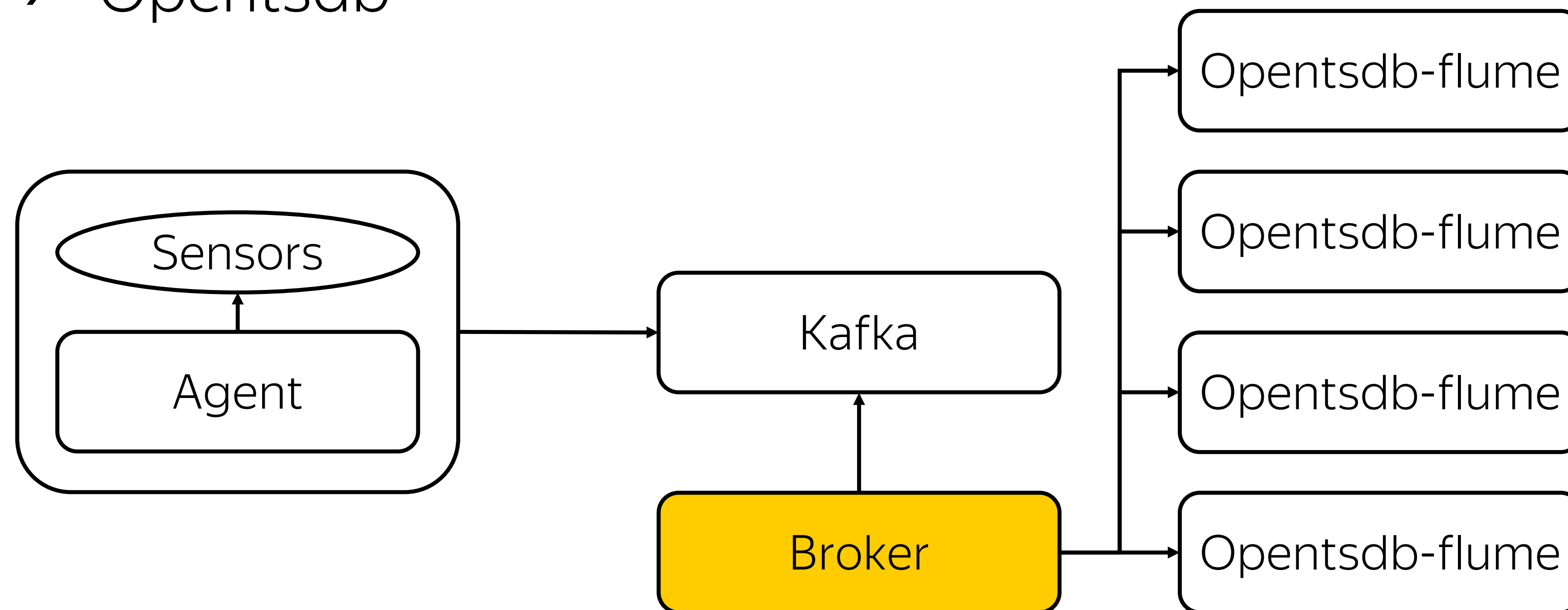
- › Apache Flume
- › Opentsdb



Yandex Tables – сбор и хранение метрик

<https://github.com/yandex/opentsdb-flume>

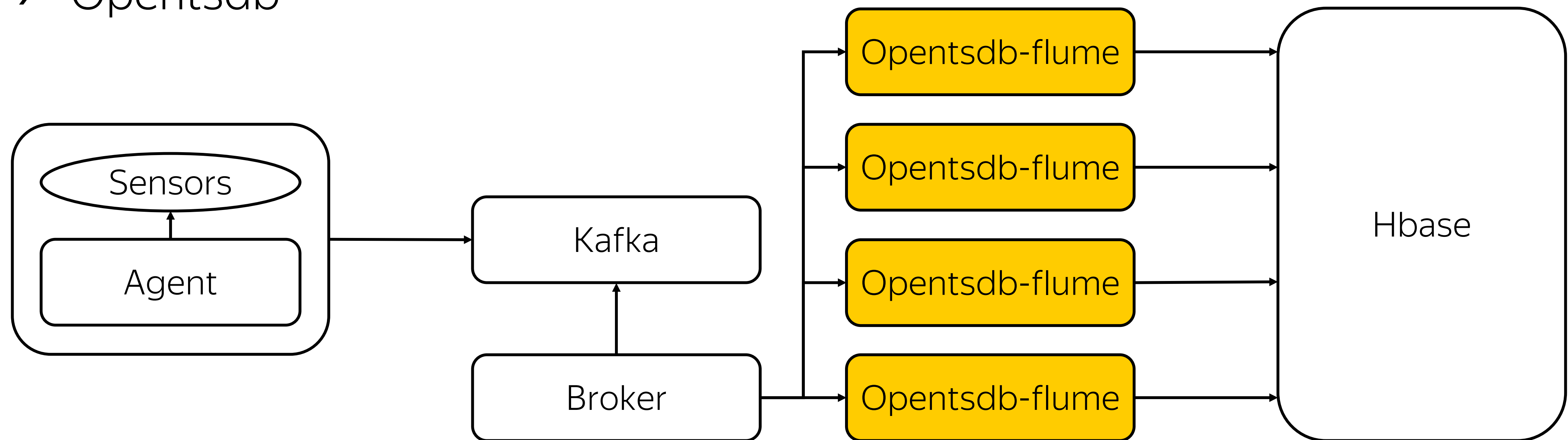
- › Apache Flume
- › Opentsdb



Yandex Tables – сбор и хранение метрик

<https://github.com/yandex/opentsdb-flume>

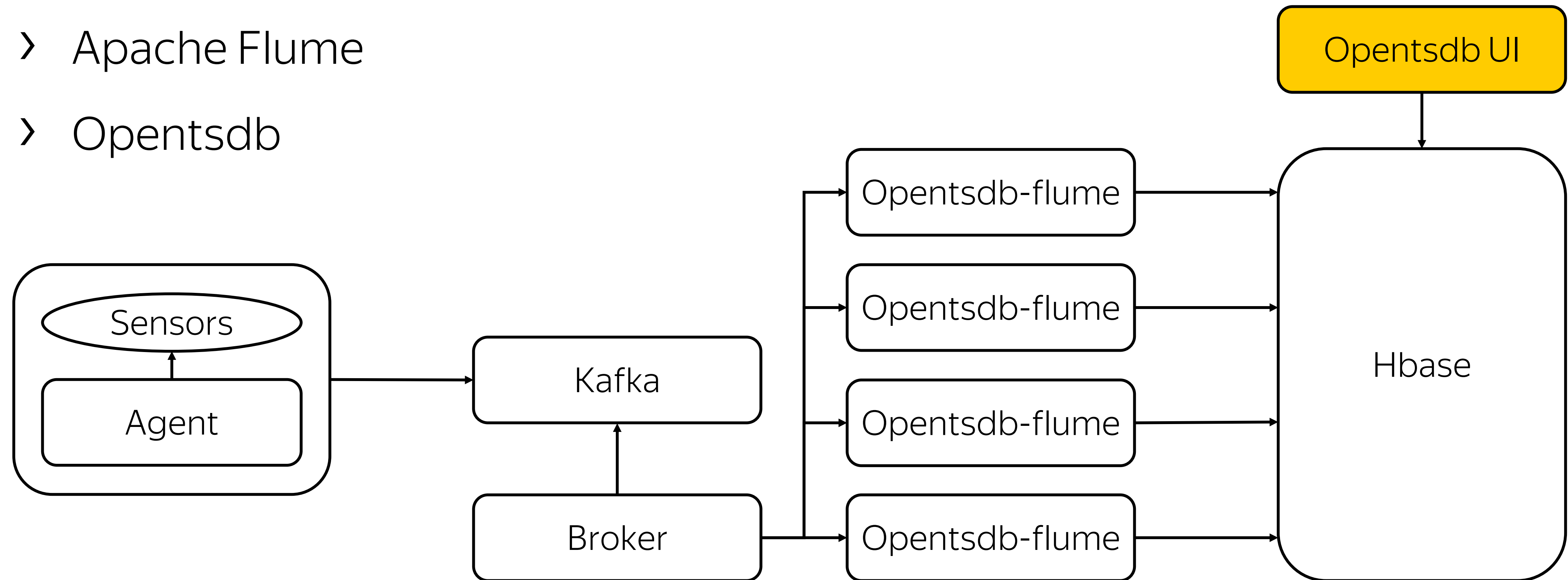
- › Apache Flume
- › Opentsdb



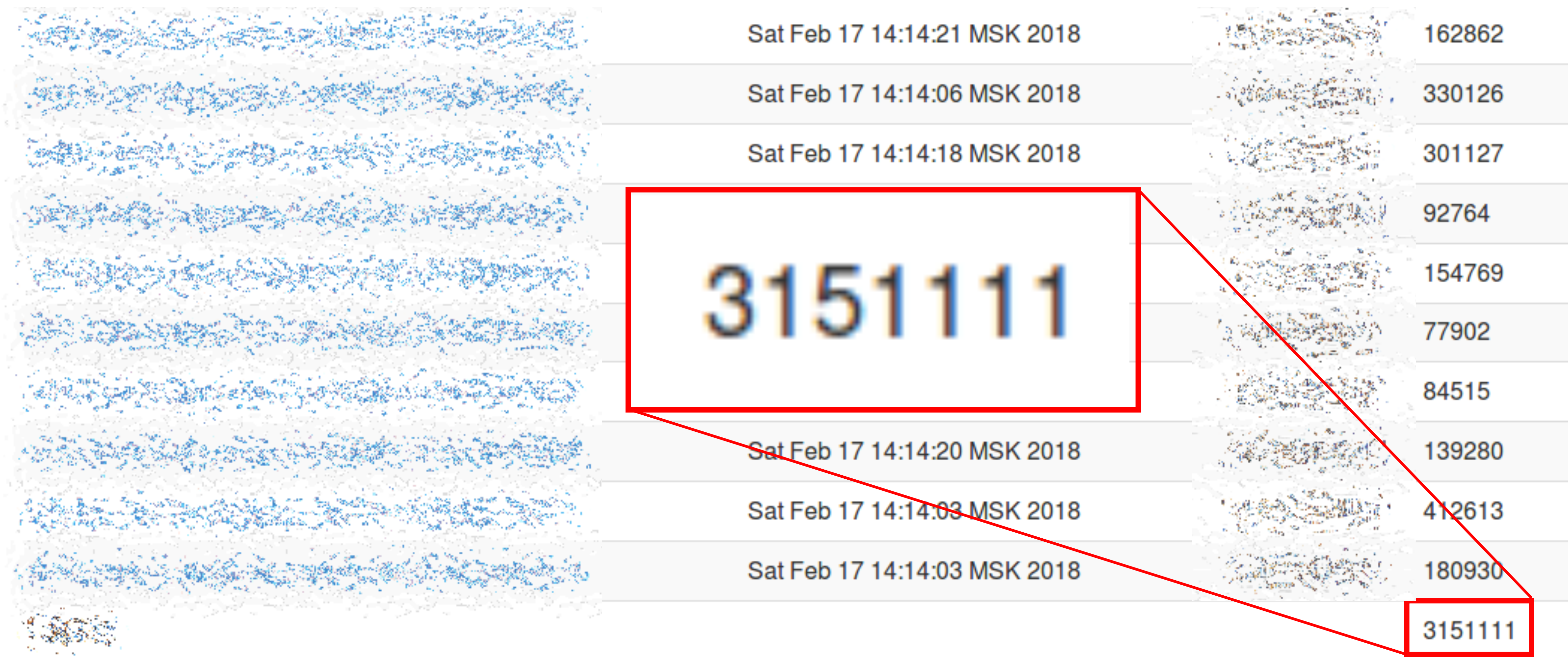
Yandex Tables – сбор и хранение метрик

<https://github.com/yandex/opentsdb-flume>

- › Apache Flume
- › Opentsdb



Opentsdb-flume + Hbase – 3.1kk RPS на запись



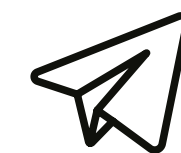
Спасибо за внимание!

Игорь Андреев

Системный администратор



iandreyev@yandex-team.ru



[@iandreyev](https://t.me/iandreyev)



8-923-184-18-75